Domain-Unified Prompt Representations for Source-Free Domain Generalization

Hongjing Niu*, Hanting Li*, Feng Zhao†, *Member, IEEE*, Bin Li, *Member, IEEE*

Abstract—Domain generalization (DG), aiming to make models work on unseen domains, is a surefire way toward general artificial intelligence. Limited by the scale and diversity of current DG datasets, it is difficult for existing methods to scale to diverse domains in open-world scenarios (e.g., science fiction and pixelate style). Therefore, the source-free domain generalization (SFDG) task is necessary and challenging. To address this issue, we propose an approach based on large-scale vision-language pretraining models (e.g., CLIP), which exploits the extensive domain information embedded in it. The proposed scheme generates diverse prompts from a domain bank that contains many more diverse domains than existing DG datasets. Furthermore, our method yields domain-unified representations from these prompts, thus being able to cope with samples from open-world domains. Extensive experiments on mainstream DG datasets, namely PACS, VLCS, OfficeHome, and DomainNet, show that the proposed method achieves competitive performance compared to state-of-the-art (SOTA) DG methods that require source domain data for training. Besides, we collect a small datasets consists of two domains to evaluate the open-world domain generalization ability of the proposed method. The source code and the dataset will be made publicly available at https://github.com/muse1998/Source-Free-Domain-Generalization

Index Terms—Dynamic facial expression recognition, expression intensity, deep learning.

✦

1 INTRODUCTION

Deep learning [1], [2] has achieved great success in many fields, especially computer vision. However, the generalization of deep learning methods is not yet good enough, which has become a constraint for wide applications. In recent years, deep learning generalizability has attracted increasing attentions and sprouted several subfields. Although these studies differ somewhat in their task settings, their ultimate goals are the same, i.e., to enabling models to cope with a wide range of possible scenarios in the open world.

Generalizability has long been an essential topic in the field of machine learning. In practical application scenarios, the ability to make reasonable predictions on unseen data is critical for machine learning models. In recent years, thanks to the advancement of algorithms and the increase in data scale, deep learning models have achieved satisfactory performance in testing scenarios that are independently and identically distributed with the training dataset. Furthermore, deep learning models could face more challenges, where the testing and training data may obey different distributions.

Domain generalization (DG) is one of the most representative tasks and has attracted considerable attentions. For image classification, the DG tasks provide datasets from multiple domains that may come from different acquisition methods or have different image styles. These data are divided into a source domain (for training) and a target domain (for testing), without overlapping between them. For example, models are trained on a photo dataset and tested on a sketch dataset.

For DG tasks, the source domain data are often an important factor indicating the difficulty of the task. When the source domain data is rich, the model learns domain-invariant features more easily and presents better generalization capabilities. In the case of PACS [3] dataset, for example, an approach using three source domains can achieve nearly double performance of the one that utilizes a single source domain. In real-world tasks, obtaining sufficiently rich source domain data is often not guaranteed. Reducing the dependence on source domain data in domain generalization is a widely applied and exciting problem, so we propose the source-free domain generalization (SFDG) task. The model is used directly to predict the target domain samples without using the source domain data for training.

To solve the SFDG problem, we apply a vision-language model for feature extraction. Compared to acquiring a huge amount of images as source domain data, representing the source domain features in the form of text is almost costless. After obtaining the textual encodings of different domains by constructing various textual prompts, we propse a domain-unified prompt representation generator (DUPRG) to integrate these text encodings and aggregate a set of domain-unified text representations. For the target domain samples from the open world, the model prediction is made based on the domain-unified text representations. DUPRG With the vision-language model, we only need to process the text encoding to obtain domain invariance in the training phase, thus bypassing the need for a large amount of source domain data and achieving source-free domain generalization.

Unlike most DG methods, our model does not require source domain data for training, which greatly extends the application scenarios of the task.

The main contributions of this work can be summarized as follows.

---

* *Equal contribution*
† *Corresponding to fzhao956@ustc.edu.cn*
H. Niu, H. Li, F. Zhao, and B. Li are affiliated with University of Science and Technology of China, Hefei 230026, China.

- The research problem of SFDG is introduced, which is more suitable for most application scenarios than the basic DG task.
- An effective SFDG method is proposed to achieve DG for visual tasks by learning domain-unified text encodings. The proposed framework has comparable or even better performance than other DG approaches using source domain training on mainstream datasets.

2 RELATED WORK

2.1 Domain Generalization

In this paper, a more challenging task, named source-free domain generalization is proposed based on the domain generalization task. In computer vision, a seminal work [4] suggests that dataset biases can lead to poor generalization performance. For example, a person classifier trained on Caltech101 [5] could obtain a very low accuracy (11.8%) on LabelMe [6]. The domain generalization issue raised attentions, and many wonderful studies have tried to address it since then. From a methodological point of view, the DG approaches can be broadly divided into domain alignment [7], [8], meta-learning [9], data augmentation [10], ensemble learning [11], self-supervised learning [12], disentangling representations [13], regularization strategies [14], etc., which has been discussed in detail in a survey [15].

Here, we discuss some nuances of the existing DG tasks from a task setting perspective. In the basic DG task, the dataset consists of samples from multiple domains [3], [16], [17], [18], [19], and in addition to the category labels, domain labels are also provided. These samples are divided into source and target domains, where the source domain is used for training and validation, while the target domain is used for testing only. Domainbed [20] provides a complete process, and many methods follow its experimental settings. Single-source DG [21] is more challenging, since its source domain has only one domain. This makes it difficult to find common domain features. Therefore, most approaches opt for less affected strategies such as data augmentation. Some attempts, like unsupervised DG [22], are made to improve the generalizability from a pretraining perspective. By replacing the commonly used ImageNet pretraining with an unsupervised pretraining approach, the model achieves better performance on downstream DG tasks. Zero-shot DG [23] modifies the category space of the target domain so that it no longer coincides with the category space of the source domain.

It can be seen that various difficult versions of previous DG tasks, including unsupervised DG and zero-shot DG, cannot completely discard the source domain.

2.2 Vision-language pretraining

In contrast to natural images, language, as a symbolic representation created by humans, is inherently rich in prior knowledge and well interpretable, which inspires researchers to exploit the rich semantic information contained in natural language to help the neural networks to learn a better visual representation [24], [25], [26], [27], [28]. These methods effectively enhance the performance of many cross-modal tasks, such as visual question answering (VQA) [29] and visual commonsense reasoning (VCR) [30].

Specially, contrastive language-image pretraining (CLIP) [31] utilizes 400 million (image, text) pairs collected from the internet to learn robust and superior image representations through contrastive learning. Recently, some methods find that CLIP is exceptional at encoding the semantic meaning of visual depictions, regardless of their styles [32], [33], which is in line with the original purpose of the domain generalization task, i.e., learning the uniform visual semantic representations across domains. Therefore, [34] introduced a novel domain generalization paradigm to better leverage various large-scale pretraining models, including CLIP. [35] devised domain prompt learning (DPL) to generate a domain-specific prompt for each image. Similarly, Zheng et al. trained a prompt adapter to produce a suitable prompt for each input image. [36] derived a tractable variational lower bound via approximating the oracle model by a pretrained model, termed as mutual information regularization with oracle (MIRO).

The above approaches require more or less training on the source domain to learn the uniform semantics between different domains. They suffer from two main problems: 1) *Limited source domain data is hard to be extended to the rich unknown domains in open-world scenarios (e.g., science fiction and pixelate style)*, and 2) *Well-developed visual semantic and rich domain information embedded in large-scale vision-language pretraining models are not fully utilized*. Therefore, we propose a novel domain-unified prompt representation generator to deal with the source-free domain generalization task in open-world scenarios.

3 METHOD

3.1 Problem Formulation

The domain generalization task aims to address the shift in data distribution between different domains by zero-shot transferring knowledge from the source domain to the unseen domain. DG methods usually train a model $f : \mathbb{R}^{H \times W} \mapsto \mathbb{R}^C$ on several source domains $D_s = \{D_1, D_2, \cdots D_S\}$ and then evaluate it on unseen target domains $D_t = \{D_1, D_2, \cdots D_T\}$. $D_K = \{(x_i, y_i)\}_{i=1}^N$ represents a dataset of the domain K, which samples N sample points from an input image space $\mathcal{X}_K \in \mathbb{R}^{H \times W}$ and a domain-shared label space $\mathcal{Y} \in \mathbb{R}^C$, where H and W are the height and width of the input images, and C is the number of the classes.

However, the existing domain generalization datasets (e.g., PACS [37], and DomainNet [38]) contain only a very limited number of domains (e.g., painting, sketch, and cartoon), which makes it difficult for networks trained on these datasets to generalize to unseen domains (e.g., mosaic and abstract) in the open-world scenarios. Therefore we first introduce a source-free domain generalization task that aims to generalize the visual representations learned by the neural networks to any unseen domain of real-world scenarios without training through the source domain data D_s (i.e., without sampling images from the source domain).
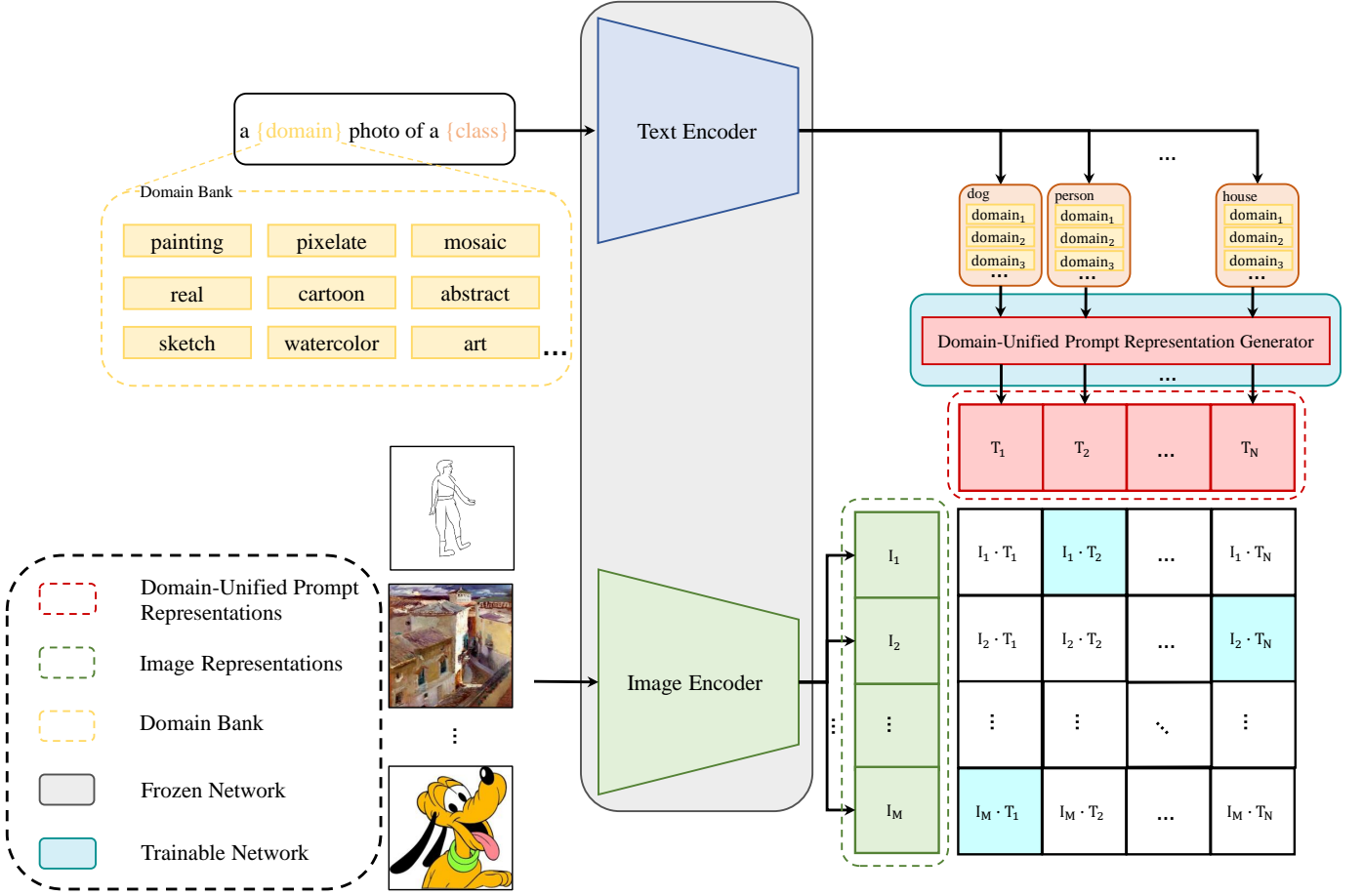
Fig. 1. Overview of our method for SFDG. First, based on the rich variety of domains in domain bank (DB), we generate M text descriptions under different domains for each class, where M is the size of the domain bank. After that, our proposed DUPRG generates a domain-unified prompt representation for each class for inference. It is worth noting that only DUPRG needs to be trained in the whole paradigm while the parameters of both text and image encoders are frozen.

3.2 CLIP for source-free domain generalization task

To achieve generalization in open-world scenarios, our model needs to consistently model different semantics under a rich set of domains. However, it is not easy to encompass the domain knowledge in the real world using only the current limited-scale DG datasets.

Recently, large-scale vision-language pretraining models [31], [39] obtain uniform and robust semantic representations [33] by contrast learning over a large number of (image, text) pairs. Especially, CLIP match the accuracy of the original ResNet50 [40] on ImageNet [41] zero-shot without needing to use any images from ImageNet. We believe that the extensive domain information embedded in CLIP is the key to solving the SFDG task.

As a vanilla approach, CLIP uses its text encoder and image encoder to embed the input prompts ("a photo of a {class}") and images into text features $T_i \in \mathbb{R}^d$ and image features $I_j \in \mathbb{R}^d$ of the same dimension d, respectively. Then CLIP gives the inference result by calculating the cosine similarity of T_i and I_j, which can be formulated as,

$$\widehat{y}_j = \arg\max_i(\langle I_j, T_i \rangle), \qquad i \in \{1, 2, \cdots, C\}. \quad (1)$$

Where $\widehat{y}_j$ is the prediction made by CLIP on the j-th image, C is the number of the classes, and $\langle \cdot, \cdot \rangle$ represents the cosine similarity between two vectors. For SFDG tasks, a naive strategy utilizing CLIP is to use a standard prompt to generalize text prompts for each category, and then finish the prediction following the above inference process.

3.3 DUPRG: a new paradigm for source-free domain generalization task

Although experiments in [35] show a good performance on several DG datasets with the standard prompt ("a photo of a {class}"). The standard prompt cannot provide an equally robust prompt for visual semantic representations of different open-world domains. For example, "a {watercolor} photo of {class}" can be a better prompt for watercolor images. To accomplish the SFDG task, we need to generate domain-unified prompt representations for each category to cope with the rich open-world domains.

Based on this viewpoint, we build a domain bank, which contains various domains that exist in the open-world scenarios. As shown in Fig. 1, according to the domain type in the domain bank, M prompts are generated for each class, which means that we can get $M \times C$ prompt representations embedded by the text encoder.
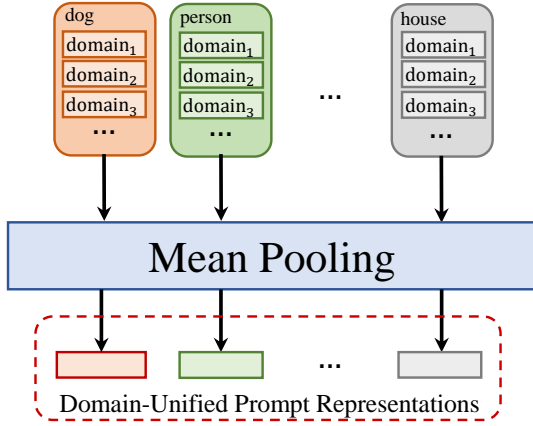
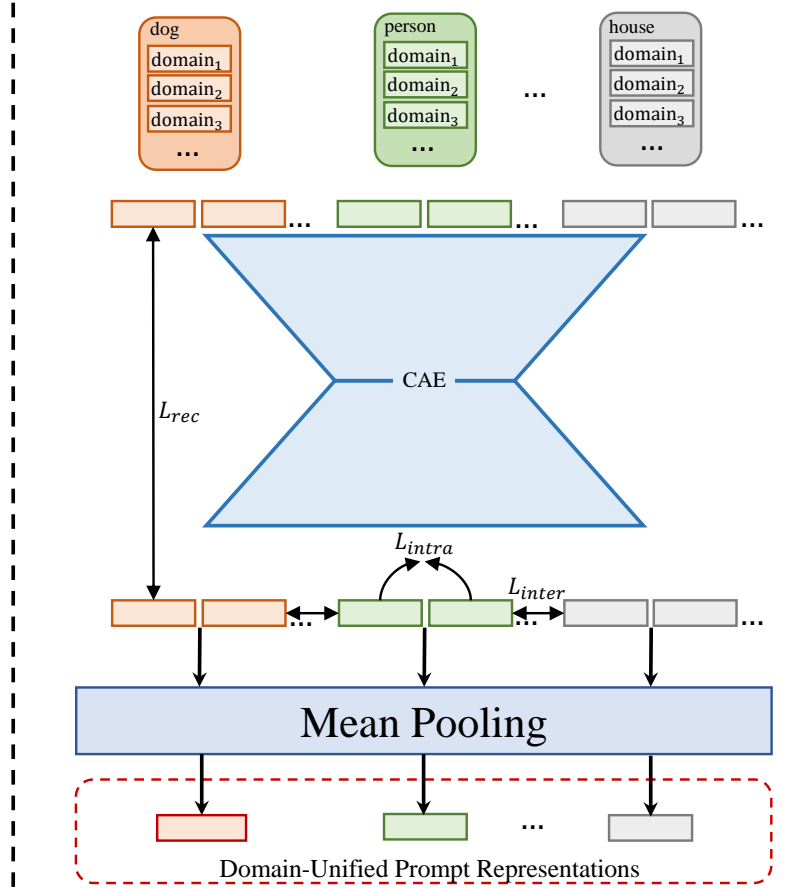Fig. 2. Two implementations of the proposed domain-unified prompt representation generator.

After getting the M prompt representations of each class, these prompt representations are aggregated by the DUPRG into a domain-unified prompt representation for final inference. We propose two specific implementations of DUPRG in Fig. 2.

Mean Pooling: As shown in Fig. 2(a), the most intuitive way to aggregate these M representations is to take the mean value as the domain-unified prompt representation. The domain-unified prompt representation of the i-th class can be defined as,

$$T_i = \frac{\sum_{j=1}^{M} T_i^j}{M}, \qquad i \in \{1, 2, \cdots, C\}, \qquad (2)$$

where T_i^j denotes the prompt representation of the i-th class in the j-th domain.

Cosine Autoencoder: As shown in Fig. 2(b), we propose a cosine autoencoder (CAE) to extract the shared part of the original M prompt representations (i.e., class semantic related part). A naive autoencoder computes the mean squared error (MSE) between the input and output as the reconstruction loss function [42]. However, the inference process of CLIP uses the cosine similarity to make the final prediction. Therefore, to preserve the semantically relevant orientation in the reconstruction process, we use cosine

similarity as the reconstruction loss of the proposed CAE. The reconstruction loss can be formulated as,

$$\mathcal{L}_{rec} = -\frac{\sum_{i=1}^{C} \sum_{j=1}^{M} \langle T_i^j, \widehat{T_i^j} \rangle}{MC}, \qquad (3)$$

where $\widehat{T_i^j}$ is the reconstruction output of CAE for T_i^j. In addition, in order to obtain a domain-unified representation, an intra-class loss is used to narrow the gap between representations of different domains, which can be computed as,

$$\mathcal{L}_{intra} = -\frac{\sum_{i=1}^{C} \sum_{j=1}^{M} \langle T_i^j, \overline{T_i} \rangle}{MC}, \qquad (4)$$

where $\overline{T_i}$ is the mean value of M prompt representations of i-th class. Finally, an inter-class loss is defined to make the cosine similarity of the representations of different classes in the same domain as low as possible (i.e., obtain a larger inter-class distance in each domain). The inter-class loss can be defined by,

$$\mathcal{L}_{inter} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{C} \sum_{k=1, j \neq k}^{C} \langle T_j^i, T_k^i \rangle}{MC(C-1)}. \qquad (5)$$

The overall loss function of CAE is defined as follows,

$$\mathcal{L}_{all} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{intra} + \lambda_2 \mathcal{L}_{inter}. \qquad (6)$$

Where λ_1 and λ_2 are two hyper-parameters controlling loss coefficients. The domain-unified prompt representation can be directly calculated by,

$$T_i = \frac{\sum_{j=1}^{M} \widehat{T_i^j}}{M}, \qquad i \in \{1, 2, \cdots, C\}, \qquad (7)$$

It is worth noting that no images are involved in the whole training process of CAE. Our method takes $M \times C$ prompt representations generated from the domain bank as the input to train the CAE. Since our method does not overfit the limited source domain image data and learns a unified prompt representation from rich domain type in the domain bank, it can be adapted to more unseen domains in richer open-world scenarios.

4 EXPERIMENTS

Our code is implemented in PyTorch (and will be open source). The datasets and experiment settings used are described below.

Datasets. We experiment on 5 benchmark datasets including PACS [3] (4 domains, 9,991 samples, 7 classes), VLCS [16] (4 domains, 10,729 samples, 5 classes), Office-Home [17] (4 domains, 15,588 samples, 65 classes), TerraIncognita [18] (4 domains, 24,778 samples, 10 classes), and DomainNet [19] (6 domains, 586,575 samples, 345 classes). These datasets are all representative and widely used in DG tasks. In a DG task, experiments are usually performed using the leave-one-out strategy. For a dataset, one of the domains is selected as the target domain at a time, and the other domains are used as the source domains. In the SFDG task proposed in this paper, one domain is also selected as the target domain each time, but the data from other domains will not be used as the source domain.

Details. We adopt the pretraining weights of CLIP and apply ViT-B/16 [43] as the backbone for the experiments without special instructions. More experiments using other backbones can be found in Appendix.2. The CAE is an autoencoder consisting of linear layers and activation functions, and the number of neurons in the hidden layer is (512, 256, 512). The optimizer is AdamW with a learning rate of 0.04. The max epochs is 1000, since the training data is scarce and CAE needs to be finely tuned to achieve a better reconstruction.

4.1 Domain Generalization

We compare our approach with some strong DG baselines including SOTA. The results are shown in Table 1. As discussed in Section 2, some of the compared methods incorporate elaborate learning algorithms and some works incorporate ensemble learning. It is noteworthy that all these methods require more or less training on the source domain. Table 1 shows that our method achieves over or near SOTA performance on most datasets. We also notice that our method does not perform well on TerraIncognita. We speculate that the possible reasons are (1) different domains are simply taken from different cameras, so the domain gaps of TerraIncognita are much smaller compared to other datasets such as PACS; (2) some categories in TerraIncognita are rare (e.g. bobcat) and have small differences

between classes (e.g., bobcat and cat), which makes source-free prediction extremely difficult.

4.2 Ablations

Hyper-parameters. Our method introduces two additional hyperparameters (i.e.,the loss weight coefficients λ_1 and λ_2 in Eq 6). We tested different combinations of λ_1 and λ_2 on PACS and VLCS datasets, which encompassed the ablation experiment (when $\lambda_1 = \lambda_2 = 0$). The results are shown in Fig. 3. We only show the experimental results on the VLCS dataset in order to provide a clearer picture of the effect of hyperparameter variation on the results. It can be seen that when both loss functions have a performance improvement. In addition, most of the hyperparameter combinations can achieve good performance, and the choice of weights does not require particularly tedious adjustment. Specially, $\lambda_1 = \lambda_2 = 0$ refers to the mean pooling strategy as shown in Fig. 2

Loss for reconstruction. Inspired by the inference process of CLIP, we choose the cosine similarity as the reconstruction loss $\mathcal{L}_{rec}$ to optimize the autoencoder. The effects of using traditional L2 loss is also given in Table 2 for a comparison.

Domain bank. The rich and diverse prompt can be considered as the raw material for constructing domain-unified prompt representations. Compared to collecting domain-specific image datasets, generating prompts in textual form can be very easy. We integrated the domains of all the used datasets into a domain bank as the defalut setting for all the experiments in this paper, which corresponds to "combined" in Table 3. In fact, the domain bank can be easily expanded or reduced. Table 3 shows the experimental results corresponding to different domain banks. On one hand, when the bank shrinks to the limit (empty set), the prompt set degenerates to the standard prompt. On the other hand, an overly redundant domain bank can increase the amount of computation while not necessarily improving performance, but may have a better performance for the open-world scenarios.

4.3 Visualizations

To further demonstrate the effect of the proposed CAE on generating domain-unified prompt representations, we visualize the input and output feature distributions of CAE by t-SNE [56] on PACS and OfficeHome datasets. As shown in Fig. 4, the input prompt representations of different domains (orange) are relatively scattered, while the output prompt representations (blue) are closely distributed. This means that out CAE filter out the domain-related part while the class-related part is preserved.

Since both strategies calculate the domain-unified prompt representations by utilizing the mean pooling operation in Fig. 2, the center of the prompt representations serves as the domain-unified prompt representations for each class. As shown in the zoomed-in part of Fig. 4(b), the center of the output features (blue) has a clear shift compared to that of the input features (orange), which is brought by is the filtering of domain-related parts. Therefore, our CAE can be used to generate better domain-unified prompt representations.

TABLE 1
Accuracy (%) on PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet. The last column is the average results of the former five columns. The best results are in **bold faces** and the second best results are underlined. MP and CAE represent the "mean pooling" and the "cosine autoencoder" strategy in Fig. 2, respectively.

Method	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg.
MMD [44]	84.7 ± 0.5	77.5 ± 0.9	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	58.8
Mixstyle [45]	85.2 ± 0.3	77.9 ± 0.5	60.4 ± 0.3	44.0 ± 0.7	34.0 ± 0.1	60.3
GroupDRO [46]	84.4 ± 0.8	76.7 ± 0.6	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	60.7
IRM [47]	83.5 ± 0.8	78.5 ± 0.5	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	61.6
Fish [48]	85.5 ± 0.3	77.8 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	42.7 ± 0.2	63.9
ERM [49]	84.2 ± 0.1	77.3 ± 0.1	67.6 ± 0.2	47.8 ± 0.6	44.0 ± 0.1	64.2
SagNet [50]	86.3 ± 0.2	77.8 ± 0.5	68.1 ± 0.1	48.6 ± 0.1	40.3 ± 0.1	64.2
SelfReg [51]	85.6 ± 0.4	77.8 ± 0.9	67.9 ± 0.7	47.0 ± 0.3	42.8 ± 0.0	64.2
CORAL [52]	86.2 ± 0.3	78.8 ± 0.6	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.1	64.5
mDSDI [53]	85.4 ± 0.4	79.0 ± 0.0	70.5 ± 0.4	50.4 ± 1.1	44.3 ± 0.2	65.9
MIRO [36]	85.4 ± 0.4	79.0 ± 0.0	70.5 ± 0.4	50.4 ± 1.1	44.3 ± 0.2	65.9
SWAD [54]	88.1 ± 0.1	79.1 ± 0.1	70.6 ± 0.2	50.0 ± 0.3	46.5 ± 0.1	66.9
DPL [35]	$\mathbf{97.3} \pm 0.2$	$\mathbf{84.3} \pm 0.4$	$\mathbf{84.2} \pm 0.2$	$\underline{52.6} \pm 0.6$	—	—
GVRT [55]	85.1 ± 0.3	79.0 ± 0.2	70.1 ± 0.1	48.0 ± 0.2	44.1 ± 0.1	65.2
SEDGE [34]	96.1 ± 0.0	82.2 ± 0.0	80.7 ± 0.2	$\mathbf{56.8} \pm 0.3$	54.7 ± 0.1	$\mathbf{74.1}$
Ours-MP (source free)	96.5 ± 0.0	83.1 ± 0.0	82.5 ± 0.0	38.3 ± 0.0	$\underline{59.2} \pm 0.0$	71.9
Ours-CAE (source free)	$\underline{97.1} \pm 0.2$	$\underline{83.9} \pm 0.5$	$\underline{83.6} \pm 0.3$	42.0 ± 1.3	$\mathbf{59.6} \pm 0.3$	$\underline{73.2}$

TABLE 2
Accuracy (%) on SFDG tasks corresponding to different loss functions. "SP" represents using standard prompt ("a photo of a **class**") for inference. The best results are in **bold faces**.

	PACS					VLCS				
	Photo	Art	Cartoon	Sketch	Avg.	VOC2007	LabelMe	Caltech101	SUN09	Avg.
SP	99.8	97.2	99.1	88.1	96.1	86.0	70.2	99.9	73.6	82.4
L2 loss	83.0	97.3	67.9	69.1	79.3	72.8	**72.8**	69.3	69.1	71.0
Cosine loss	**99.9**	**97.8**	**99.2**	**91.4**	**97.1**	**87.0**	72.5	**99.9**	**76.1**	**83.9**

TABLE 3
Accuracy (%) on SFDG tasks corresponding to different domain banks. An "empty" bank means using standard prompt only; "task" means to form the bank with domain type of the corresponding dataset; "combined" represents a summary of the domains of several datasets; and "expanded" adds some additional propmts based on "combined". The domains of VLCS and TerraIncognita do not correspond to a specific style, so no specific prompt is provided. The best results are in **bold faces** and the second best results are underlined.

Bank	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg.
Empty	96.1	82.4	81.6	36.9	56.7	70.7
Task	96.7	82.4	**83.8**	36.9	57.2	71.4
Combined	**97.1**	**83.9**	$\underline{83.6}$	**42.0**	**59.6**	**73.2**
Expanded	$\underline{96.8}$	$\underline{82.9}$	83.3	$\underline{40.8}$	$\underline{59.1}$	$\underline{72.6}$

TABLE 4
Accuracy (%) on domains in open-world scenarios. "X", "G", "P", "A", "C", and "S" represent the pixelate and geometric, photo, art, cartoon, and sketch style, respectively. The best results are in **bold faces**.

Multi-Source Methods

	P A C		P A S		P C S		A C S	
	X	G	X	G	X	G	X	G
ERM [49]	51.5	61.5	48.2	65.1	57.9	58.9	53.8	62.1
GroupDRO [46]	52.2	61.3	51.6	62.9	60.4	59.7	55.3	64.1
CORAL [52]	46.1	57.1	45.0	62.2	51.1	57.2	49.2	63.5
DPL [35]	94.9	90.8	91.9	91.3	96.0	93.8	93.7	92.9

Source-Free Methods

	X		G	
Ours (CAE)	97.1		95.1	

4.4 Open-world Domain Generalization

To demonstrate the advantages of our approach on the domain generalization task in open-world scenarios, we collected data of two domains (i.e., pixelate style and geometric style) that hardly appear in the current DG datasets from internet as a supplement to the PACS dataset (share the same categories with the PACS dataset). More details about the datasets can be found in Appendix.1. As shown in Table 4, our method that need no source domain data for training significantly outperforms recent DG methods by a large margin.
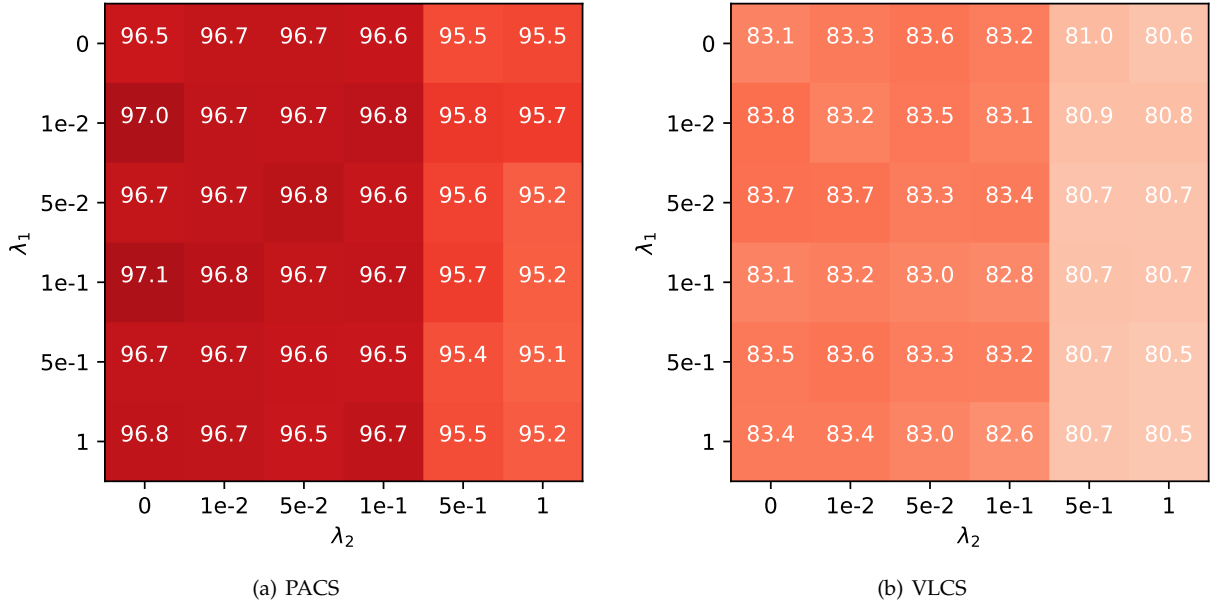
(a) PACS

(b) VLCS

Fig. 3. Average accuracy on PACS and VLCS datasets with different combinations of weighting coefficients $[\lambda_1, \lambda_2]$.
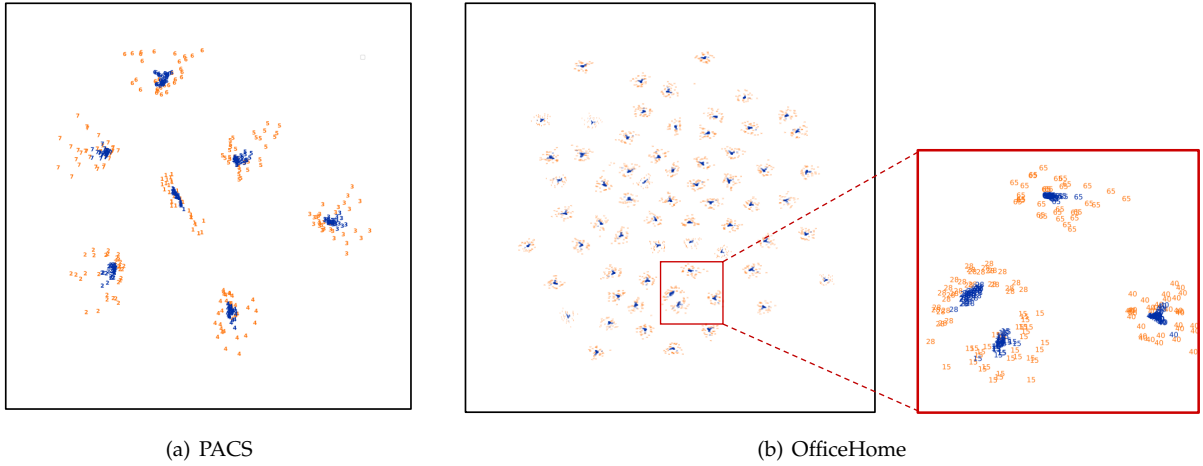

(a) PACS

(b) OfficeHome

Fig. 4. Visualization of the input and output prompt feature distribution of CAE by t-SNE. The numbers represent the categories. The orange and blue points refer to the input and output prompt representations of CAE, respectively. The points with the same color and number stand for the prompt representations generated by various domain types from the domain bank.

5 CONCLUSION

In this work, we develop a more challenging version of the domain generalization problem that requires good performance in the target domain without using the source domain data. This paper also proposes a feasible solution to the source-free domain generalization problem. By transforming the need for image feature consistency into constructing domain-unified prompt representations through a vision-language model, we bypass the need for source domain data. Experimental results demonstrate that our method exhibits excellent performance on most of the tested datasets without using source domain data, achieving a competitive performance compared to other approaches trained with source domain data. More importantly, we substantially outperformed current SOTA DG methods on open-world domains without training with multi-source domain data.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] 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.

[4] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, 2011, pp. 1521–1528.

[5] 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 *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 178–178.

[6] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1, pp. 157–173, 2008.

[7] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28, no. 1. PMLR, 17–19 Jun 2013, pp. 10–18.

[8] 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)*, September 2018.

[9] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc.

[10] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *European conference on computer vision*. Springer, 2020, pp. 561–578.

[11] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[12] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *European Conference on Computer Vision*. Springer, 2012, pp. 158–171.

[14] H. Wang, Z. He, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," in *International Conference on Learning Representations*, 2019.

[15] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] 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 (ICCV)*, December 2013.

[17] 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 (CVPR)*, July 2017.

[18] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] 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 (ICCV)*, October 2019.

[20] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021.

[21] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 834–843.

[22] X. Zhang, L. Zhou, R. Xu, P. Cui, Z. Shen, and H. Liu, "Towards unsupervised domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4910–4920.

[23] U. Maniyar, A. A. Deshmukh, U. Dogan, V. N. Balasubramanian *et al.*, "Zero shot domain generalization," *arXiv preprint arXiv:2008.07443*, 2020.

[24] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pretraining of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[25] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[26] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 336–11 344.

[27] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[28] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[29] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[30] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.

[31] 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.

[32] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir, "Clipasso: Semantically-aware object sketching," *arXiv preprint arXiv:2202.05822*, 2022.

[33] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah, "Multimodal neurons in artificial neural networks," *Distill*, vol. 6, no. 3, p. e30, 2021.

[34] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li, "Domain generalization using pretrained models without fine-tuning," *arXiv preprint arXiv:2203.04600*, 2022.

[35] X. Zhang, Y. Iwasawa, Y. Matsuo, and S. S. Gu, "Amortized prompt: Lightweight fine-tuning for clip in domain generalization," *arXiv preprint arXiv:2111.12853*, 2021.

[36] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," *arXiv preprint arXiv:2203.10789*, 2022.

[37] 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*, 2017, pp. 5542–5550.

[38] 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.

[39] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv preprint arXiv:2201.12086*, 2022.

[40] 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.

[41] 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.

[42] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[43] 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," in *International Conference on Learning Representations*, 2021.

[44] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2021.

[46] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *International Conference on Learning Representations*, 2020.

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

[48] Y. Shi, J. Seely, P. Torr, S. N, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in *International Conference on Learning Representations*, 2022.

[49] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021.

[50] 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 (CVPR)*, June 2021, pp. 8690–8699.

[51] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9619–9628.

[52] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[53] M.-H. Bui, T. Tran, A. Tran, and D. Phung, "Exploiting domain-specific features to enhance domain generalization," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Curran Associates, Inc., pp. 21 189–21 201.

[54] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Curran Associates, Inc., pp. 22 405–22 418.

[55] S. Min, N. Park, S. Kim, S. Park, and J. Kim, "Grounding visual representations with texts for domain generalization," *arXiv*, vol. abs/2207.10285, 2022.

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

APPENDIX

.1 Openset Datasets

As declared in Section 4.4, we collect a supplementary for PACS datasets to evaluate the effectiveness of our method on the domain generalization task in open-world scenarios. Specifically, we collect 1365 pixelate style images and 1202 geometric style images from the internet. As shown in Fig.5 and Fig.6, some samples in the datasets are given. The datasets will be released together with our source code.

.2 Backbones

In the proposed framework, the only part that needs to be trained is the proposed CAE (domain-unified prompt representation generator), while the parameters of the text and image encoders are fixed. These encoders are replaced to verify the generalizability of the proposed approach for different model structures. We implemented ResNet [2] and ViT [43] and generate prompt representations with different strategies, **"SP"**: using standard prompt ("a photo of a {class}") to generate the only representation for each category; **"MP"**: the prompts generated by different domains are aggregated by the mean pooling operation; **"CAE"**: using CAE to generate the unified representations. The results are shown in Table 5. Regardless of the model structure, using CAE to generate unified-representations is always the best for both ViT-based backbones and is comparable with the mean pooling strategy for ResNet50, which demonstrate the effectiveness of the proposed CAE.
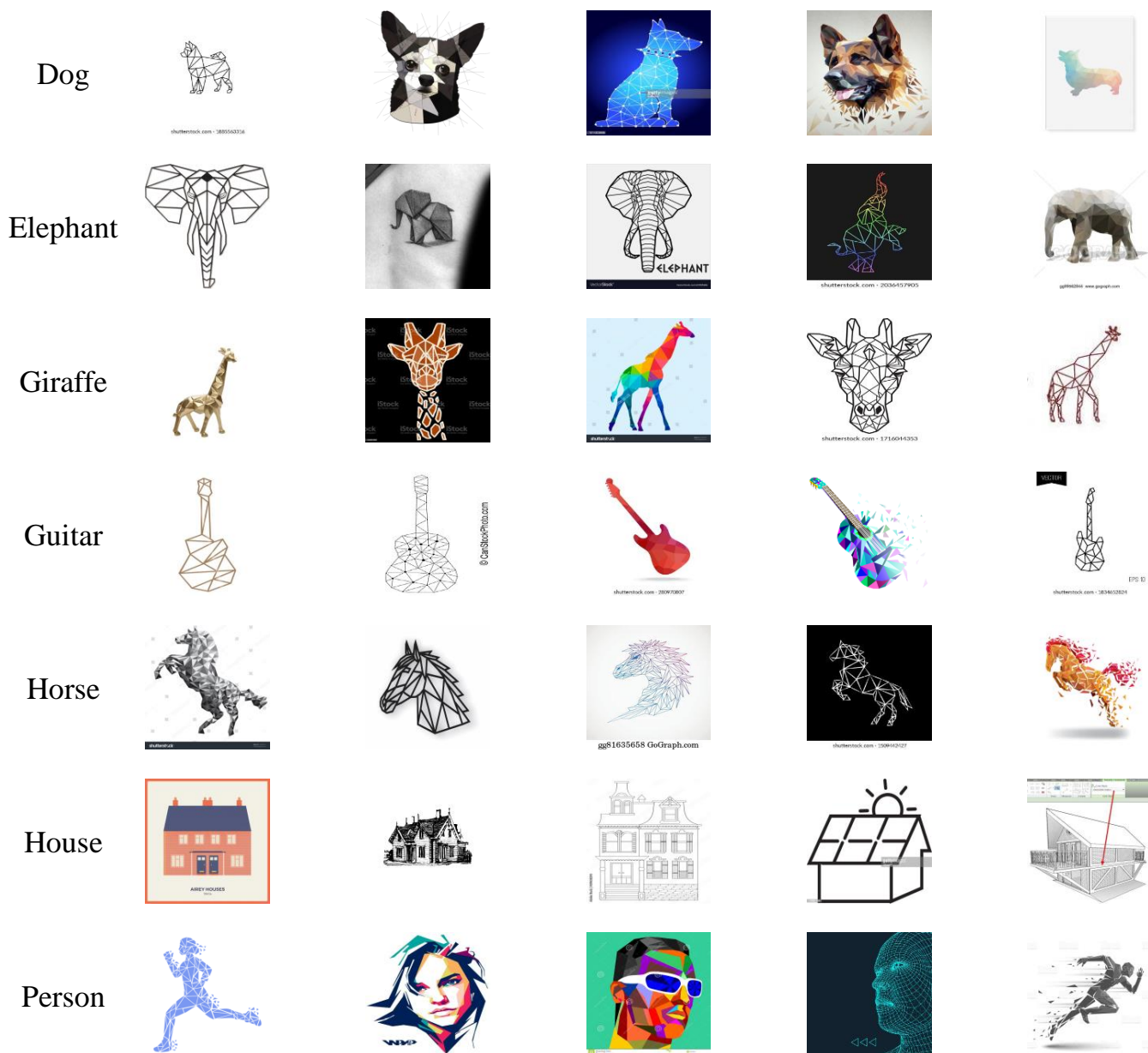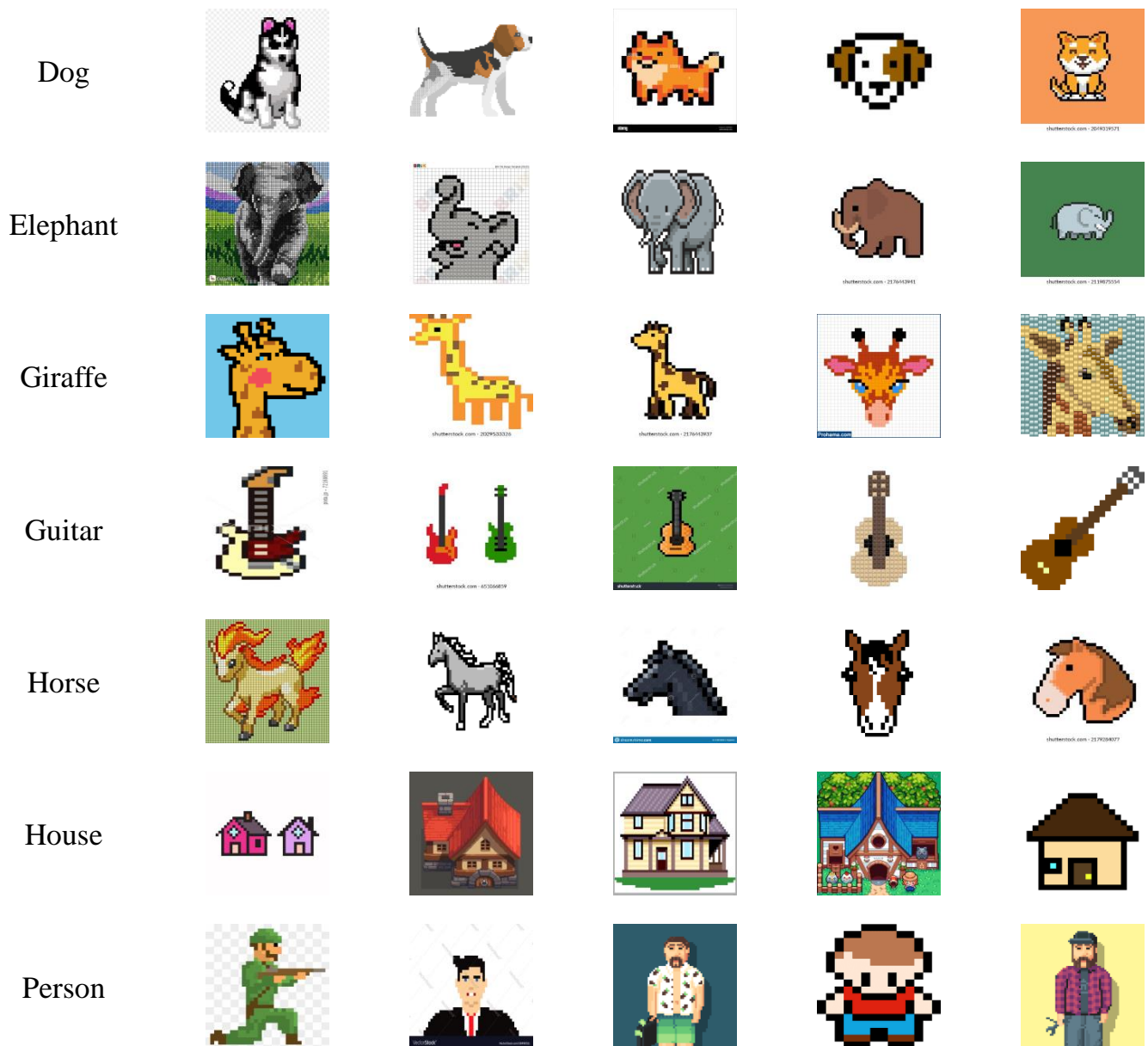
Fig. 5. Geometric style images in our dataset

Fig. 6. Pixelate style images in our dataset

TABLE 5
Accuracy (%) on SFDG tasks corresponding to different encoder backbones. Bold denotes the best.

Backbone	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg.
ResNet50 (SP)	91.1	80.8	71.3	17.8	45.6	61.3
ResNet50 (MP)	93.0	**82.5**	**75.3**	24.6	48.1	**64.7**
ResNet50 (CAE)	**93.1**	81.2	73.0	**25.2**	**48.4**	64.2
ViT-B/16 (SP)	96.1	82.4	81.6	36.9	56.7	70.7
ViT-B/16 (MP)	96.5	83.1	82.5	38.3	59.2	71.9
ViT-B/16 (CAE)	**97.1**	**83.9**	**83.6**	**42.0**	**59.6**	**73.2**
ViT-B/32 (SP)	94.7	81.3	78.8	23.2	53.2	66.2
ViT-B/32 (MP)	94.7	81.1	79.6	26.4	55.0	67.4
ViT-B/32 (CAE)	**95.1**	**81.9**	**80.3**	**39.9**	**55.7**	**70.6**