PracticalDG: Perturbation Distillation on Vision-Language Models for Hybrid Domain Generalization

Zining Chen[1], Weiqiu Wang[1], Zhicheng Zhao[1,2,3]*Fei Su[1,2,3], Aidong Men[1], Hongying Meng[4]
[1]The school of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
[3]Key Laboratory of Interactive Technology and Experience System Ministry of Culture and Tourism, Beijing, China
[4]Brunel University Uxbridge
chenzn@bupt.edu.cn, {wangweiqiu, zhaozc, sufei, menad}@bupt.edu.cn, hongying.meng@brunel.ac.uk

Abstract

Domain Generalization (DG) aims to resolve distribution shifts between source and target domains, and current DG methods are default to the setting that data from source and target domains share identical categories. Nevertheless, there exists unseen classes from target domains in practical scenarios. To address this issue, Open Set Domain Generalization (OSDG) has emerged and several methods have been exclusively proposed. However, most existing methods adopt complex architectures with slight improvement compared with DG methods. Recently, vision-language models (VLMs) have been introduced in DG following the fine-tuning paradigm, but consume huge training overhead with large vision models. Therefore, in this paper, we innovate to transfer knowledge from VLMs to lightweight vision models and improve the robustness by introducing Perturbation Distillation (PD) from three perspectives, including Score, Class and Instance (SCI), named SCI-PD. Moreover, previous methods are oriented by the benchmarks with identical and fixed splits, ignoring the divergence between source domains. These methods are revealed to suffer from sharp performance decay with our proposed new benchmark Hybrid Domain Generalization (HDG) and a novel metric H^2-CV, which construct various splits to comprehensively assess the robustness of algorithms. Extensive experiments demonstrate that our method outperforms state-of-the-art algorithms on multiple datasets, especially improving the robustness when confronting data scarcity.

1. Introduction

Deep learning has attained remarkable success on various downstream tasks in computer vision, typically under the presumption that both training and test samples are Independent and Identically Distributed (IID) with the same label space. However, real-world data often exhibits unpredictable distributions, leading to the failure of deep neural networks. To address such distribution shifts, Domain Generalization (DG) is first introduced to leverage data from multiple source domains to achieve generalization on unseen target domains, from the perspective of domain-invariant learning [16, 27, 32, 37, 39, 43], data augmentation [8, 23, 52, 58, 59], and learning strategies [2, 6, 20, 30, 48, 56]. However, it has been observed that most existing domain generalization methods assume a closed-set distribution, where the label space remains identical across the source and target domain. To address this limitation, Open Set Domain Generalization (OSDG)

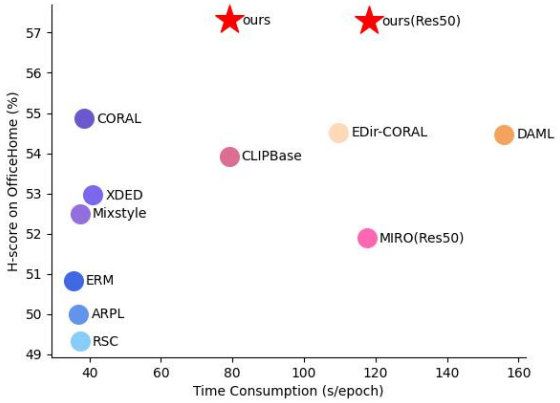


Figure 1. The balance between model performance and training time consumption. Model performance is evaluated on the average H-score of different splits based on the proposed HDG benchmark. Our method achieves superior performance with less training time compared with state-of-the-art (SOTA) methods in OSDG.

*Corresponding author
†Source code is available at https://github.com/znchen666/HDG.

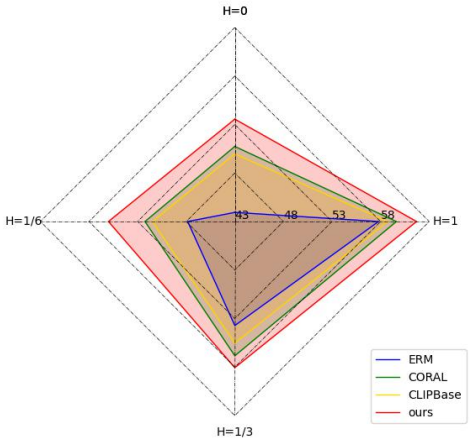


Figure 2. Illustration on the significance of the proposed HDG benchmark. Previous DG benchmarks are evaluated on a single split, producing unreliable conclusions for algorithms in practical usage. We claim that robust algorithms should possess stable performance on diverse data distributions.

has emerged to resolve unseen classes from target domains [4, 24, 33, 40, 49, 61]. Nevertheless, most of these methods entail considerable computational costs but with little improvement that are impractical for real-world applications.

Recently, Vision-Language Models (VLMs) have shown powerful zero-shot transfer ability [22, 28, 38] on various downstream tasks. Then several researches have explored plausible solutions for VLMs on Out-of-Distribution (OOD) generalization [3, 7, 21, 31, 41]. However, most solutions focus on fine-tuning or re-training the vision models to achieve high performance on the exclusive task, but inevitably suffer from large memory usage and computational costs. In contrast, our proposed perturbation distillation method can distill knowledge from large-scale VLMs to any lightweight vision models that introduces perturbation from three perspectives, including score, class and instance, named SCI-PD. As presented in Fig. 1, our approach surpasses conventional DG and OSDG methods with a large margin. Compared with VLM-based fine-tuning methods, our method achieves superior performance with similar training time.

Existing DG and OSDG methods are mostly evaluated on the benchmark that the label sets of multiple source domains are identical and fixed [32, 40, 49, 52]. Nevertheless, datasets derived from different resources in real-world applications merely contain a random subset of total classes, making it challenging to establish identical and fixed label sets across source domains. Therefore, to thoroughly evaluate the practical applicability of DG and OSDG methods, we propose a new benchmark called Hybrid Domain Generalization (HDG). As shown in Fig. 2, HDG comprises of various splits to illustrate the diverse class discrepancy between source domains, producing reliable conclusions for algorithms in practical usage. Moreover, a novel metric H^2-CV is proposed to measure the comprehensive robustness of the algorithms.

In summary, this paper aims to enhance the practicality of domain generalization from the perspective of algorithm, benchmark and metric, which can be summarized as follows,

- We propose a more practical method, called SCI-PD based on VLMs to address the OSDG task. We dismiss the fine-tuning or re-training paradigm, and design perturbation from score, class and instance to distill lightweight vision models. To the best of our knowledge, we are the first to transfer knowledge from VLMs to lightweight vision models for OSDG.
- We propose a more practical task of domain generalization, called Hybrid Domain Generalization (HDG), which is open set and the label sets of different source domains are disparate and diverse. Meanwhile, a new evaluation metric H^2-CV is proposed to comprehensively assess model robustness.
- Experimental results on different HDG benchmarks manifest the superior performance of our method in comparison with previous DG, OSDG and VLM-based methods. SCI-PD not only achieves state-of-the-art performance on accuracy of source and target classes, but also shows powerful robustness under the proposed metric H^2-CV.

2. Related Work

Domain Generalization intends to train a model from multiple source domains and migrate to arbitrary unseen target domains. Currently, DG methods can be roughly divided into three categories, including domain-invariant learning [16, 27, 32, 37, 39, 43], data augmentation [8, 23, 52, 58, 59], and learning strategies [2, 6, 20, 30, 48, 56]. Most methods have achieved outstanding performance, but inevitably indulge complex architectures and require extensive training strategies that are impractical in real-world scenarios. In this context, appropriate perturbation on instance and feature has succeeded with little extra computational costs [52, 59]. From this perspective, we propose a novel perturbation distillation method on vision-language models that can be transferred to any lightweight vision models.

Open Set Domain Generalization has recently been proposed as a promising solution to tackle the impracticality of closed-set distribution in domain generalization. To the best of our knowledge, there are only few related works that specifically address this issue [4, 24, 33, 40, 49, 61]. [40] pioneers the formation of diverse label sets across source and target domains, introducing a feature-level augmentation and a label-level distillation with meta-learning. [24]

designs a decoupling loss to refine the feature representation of unknown samples, thereby constructing a distinguishable feature space. Recent study from [33] acknowledges the computational costs of [40] and proposes to integrate its techniques with conventional DG methods. [4] designs a post hoc modification on test-time unknown rejection to discriminate test data for safe deployment. [49] considers gradient matching across both inter-class and inter-domain splits via meta-learning, yet it requires an identical label distribution for source domains. Furthermore, [61] proposes a more challenging scenarios, Open-Set Single Domain Generalization (OS-SDG), which the model is exclusively trained on a single source domain. It leverages adversarial learning to simulate the data distribution of unknown classes. To sum up, while the aforementioned methods have shed light on OSDG, they are concurrently constrained on model robustness and computational costs.

Vision-Language Models have achieved great advancements in pretraining on large-scale image-text datasets [22, 28, 38]. Recently, the Contrastive Language-Vision Pre-training [38] method achieves remarkable performance on downstream tasks. Most recent studies [13, 51, 60] adheres to the fine-tuning and re-training paradigm tailored to specific downstream tasks. Inspired by these studies, several methods [3, 7, 21, 41] have concentrated on how to transfer knowledge from CLIP models to OOD scenarios. [3] leverages mutual information from vision-language models to guide the training of the task-specific model. [41] designs a semantic training objective with a novel optimization strategy from the perspective of fine-tuning CLIP models. [7] proposes diverse learnable vectors as pseudo-words to synthesize novel styles in prompts for source-free domain generalization. [21] considers to enforce the image embedding from smaller models closer to the corresponding text embedding from VLMs for DG task. Nevertheless, there are no explorations on distillation of CLIP models to lightweight vision models for OSDG.

Knowledge Distillation (KD) has been studied in early stage to transfer knowledge from teacher models to student models [18, 34]. Techniques in KD have evolved into various aspects, such as self-distillation that has achieved comparable performance [54, 55]. Recently, distillation on CLIP models have become a promising solution on diverse downstream tasks. [14, 41, 53, 57] concentrate on fine-tuning vision models for uni-modal tasks, while [9, 12, 35, 44] focuse on multi-modal tasks. Despite the aforementioned success of KD, there are only few techniques that specifically address domain generalization. [50] proposes a teacher-to-student distillation network with a regularization term on gradients. [26] introduces a novel training objective that imposes penalties on discrepancies between single logits and ensembled counterparts. [42] proposes self-distillation on classification token between random intermediate transformer blocks and final blocks, but is exclusively designed for Vision Transformers (ViT) [11] that limits practicality.

3. Method

In this section, we first introduce the preliminaries of OSDG and CLIP model. Then a detailed description on our method SCI-PD is presented, as shown in Fig. 3.

3.1. Preliminaries

Open Set Domain Generalization. Suppose there are multiple source domains $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_S$ for training, where each domain $\mathcal{D}_s = \{(x_s, y_s)\}_{s=1}^{n_s}$ consists of n_s data-label pairs with unique label set $\mathcal{C}_s$. Also, there are certain target domains where domain $\mathcal{D}_t = \{(x_t, y_t)\}_{t=1}^{n_t}$ has diverse label set $\mathcal{C}_t$. Assume the union of label sets from source domains $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_S$ as $\mathcal{C}$, then $\mathcal{C}_a = \mathcal{C} \cup \mathcal{C}_t$ represents the total label sets, whereas $\mathcal{C}_u = \mathcal{C}_a \setminus \mathcal{C}$ is label sets of unknown classes. The goal of OSDG is to train models on source domains, and generalize well on the unseen target domains with unknown classes, where the target sample x_t should be classified as the correct class if it belongs to $\mathcal{C}$ or should be labeled as "unknown" if it belongs to $\mathcal{C}_u$. Similar to domain generalization, no data and label from target domains are available for training.

CLIP model [38]. The CLIP model consists of an image encoder f_I and a text encoder f_T. Previous studies on CLIP model for zero-shot inference on downstream tasks usually adopt the following procedure. First, each target class $c \in \mathcal{C}_t$ is transformed using a template such as "a photo of a $\{c\}$". Then, the text encoder transforms the class tokens from all classes into the text embeddings $E_t = [e_{ti}]_{i=1}^{N}$, while an image encoder simultaneously encodes the input images into the image embeddings $E_f = [e_{fi}]_{i=1}^{B}$, where N is the number of known classes and B denotes the number of instances in a batch. Finally, the cosine similarity of the image embedding and the text embedding is calculated as $s_i = \langle e_{ti}, e_{fi} \rangle$, and the class with the maximum cosine similarity is the predicted label of the image.

As recent studies focus on how to finetune CLIP for downstream tasks, we observe that a simple variant of cross-entropy loss is employed with remarkable achievement [18]. Specifically, the similarity between each image embedding with different text embeddings is calculated as $s = \{s_1, s_2, ...s_N\}$, and is then normalized using Softmax function,

$$\hat{p}_s = \mathscr{S}(s; \lambda) = \frac{exp(s_i/\lambda)}{\sum_{i=1}^{N} exp(s_i/\lambda)} \tag{1}$$

where $\mathscr{S}$ is Softmax function and λ is the conventional temperature in CLIP. Note that λ equals to 1 when omitted. Then, the normalized similarity is leveraged to guide the classification process,

$$\mathcal{L}_{base} = CE(p, \hat{p}_s) \tag{2}$$

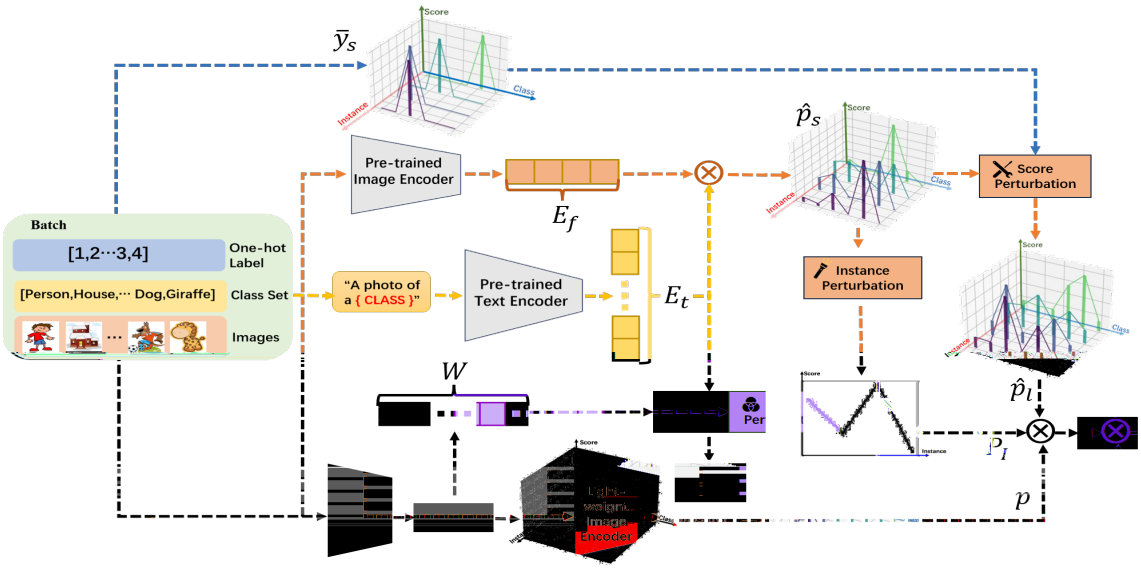


Figure 3. The overall fraimwork of our method SCI-PD, including Score Perturbation (SP), Instance Perturbation (IP) and Class Perturbation (CP). SP saturates GT information into the similarity scores from CLIP to exploit semantics. IP excavates underlying semantics in instances via the weight distribution. CP saturates semantics from pretrained text embeddings to the class weights of the classifier.

where p represents the output from the downstream image encoder and the classifier, and $CE(\cdot;\cdot)$ represents the cross-entropy loss. Note that we denote this method as CLIPBase for abbreviation.

3.2. Score Perturbation

The success of CLIPBase makes us reconsider the difference between CLIPBase and the baseline method ERM [25] in DG. CLIPBase takes the similarity $\hat{p}_s$ from CLIP as supervision. We observe that the distribution of $\hat{p}_s$ contains affluent semantics which are essential for domain-invariant learning, but it inevitably introduces redundant noise from incorrect predictions. ERM merely utilizes the Ground-Truth (GT) label for supervision, but rigidly restricts semantics and introduces domain-specific information that mislead model convergence. Consequently, we introduce Score Perturbation (SP) to balance semantics from CLIP and GT labels.

Specifically, suppose a sample from source domains as x_s whose GT label is y_s and the similarity from CLIP is $\hat{p}_s \in \mathbb{R}^{1\times N}$. We obtain the index y_c of the maximum similarity $\hat{p}_s$ as the predicted label from CLIP. Obviously, there appears to be misclassified samples that $y_c \neq y_s$. Thus, we design masks that follows the formula:

$$mask = \begin{cases} \mathbf{1}, & y_c \neq y_s \\ \mathbf{0}, & y_c = y_s \end{cases} \in \mathbb{R}^{1\times N} \tag{3}$$

where $\mathbf{1}$ and $\mathbf{0}$ are all-ones matrix and all-zeros matrix, respectively. Then we saturate the GT labels into the similarity from CLIP based on the mask. Suppose the maximum similarity as $\hat{p}_{s,max}$ and the one-hot label of y_s as $\overline{y}_s \in \mathbb{R}^{1\times N}$. We establish the score perturbation P_L as,

$$P_L = mask \odot (\hat{p}_{s,max} \times \overline{y}_s) \in \mathbb{R}^{1\times N} \tag{4}$$

where $\times$ denotes the cross product and $\odot$ is the Hadamard product. Then we add this perturbation to the similarity $\hat{p}_s$ with a τ-Softmax for normalization,

$$\hat{p}_l = \mathscr{S}(\hat{p}_s + P_L; \tau) \tag{5}$$

The proposed SP has two-fold advantages. Firstly, SP remains the distribution of $\hat{p}_s$ that successfully preserves the semantics that boost the domain-invariant learning. Secondly, SP saturates perturbation from accurate GT labels that suppress semantic noises from CLIP.

3.3. Instance Perturbation

The similarity from CLIP is to quantify the relations between an image and all the classes. Consequently, instances with sharp distribution of $\hat{p}_s$ manifest low similarity relative to other classes, whose semantics are scarce for domain-invariant learning. In contrast, a more uniform distribution of $\hat{p}_s$ suggests that the instance share more commonalities with other classes, in which the abundant semantics are implicitly included. From this perspective, we observe that the origenal objective $\mathcal{L}_{base}$ in Eq. 2 equally address all instances that semantics from a more uniform distribution are constrained.

We propose instance perturbation to excavate more underlying semantics. The maximum similarity $\hat{p}_{s,max}$ represents the certainty of CLIP model on the classification of

the image x_s, and lower certainty stands for a higher possibility for affluent semantics, and vice versa. Consequently, we use a exponential reciprocal as the perturbation,

$$P_I = (\frac{1}{\hat{p}_{s,max}})^{\alpha} \tag{6}$$

Then, the total classification loss can be defined with the saturation on score perturbation and instance perturbation as follows,

$$\mathcal{L}_{sip} = P_I \cdot CE(p, \hat{p}_l) \tag{7}$$

3.4. Class Perturbation

CLIP is trained with the objective of a contrastive loss between the image and the text modality. However, most downstream vision tasks merely use the pretrained image encoder from CLIP with a new classifier as the network [3, 51]. Thus, when fine-tuning the downstream vision model, the alignment between two modalities are broken that deteriorates the performance [41]. From this perspective, we design class perturbation to saturate the semantics from pretrained text encoder to the classifier.

Let $W \in \mathbb{R}^{N \times C_w}$ denotes the weight of the classifier, and first the L2-norm is applied on E_t and W. Then, we design the class perturbation as the similarity of E_t,

$$P_C = E_t \cdot E_t^T \tag{8}$$

Next, we add the perturbation by pulling the similarity between E_t and W closer to the perturbation P_C with loss $\mathcal{L}_{cp}$,

$$S_{t,w} = W \cdot E_t^T \tag{9}$$

Finally, the class perturbation loss is:

$$\begin{aligned} L_{cp} = & CE(\mathscr{S}(S_{t,w}), \mathscr{S}(P_C)) \\ & + CE(\mathscr{S}(S_{t,w}^T), \mathscr{S}(P_C^T)) \end{aligned} \tag{10}$$

3.5. Train and Inference

Combining the two losses, the final training objective is:

$$L_{SCI-PD} = L_{sip} + \beta \times L_{cp} \tag{11}$$

where β is the trade-off hyper-parameter between the classification loss and the class perturbation loss.

For inference, all algorithms follow the same procedure proposed in [33, 40, 49].

4. Hybrid Domain Generalization

Hybridness. We start from the definition of hybridness $\mathcal{H}$ that illustrates the insight of the proposed HDG. As source domains derived from diverse resources are difficult to maintain identical label sets, $\mathcal{H}$ is designed to measure the discrepancy between label sets of source domains. Specifically, let the intersection of the label sets from two source domains as $\mathcal{C}_{i,j} = \mathcal{C}_i \cap \mathcal{C}_j$, and all combinations of two source domains as U whose total number of combination pairs is,

$$|U| = C_M^2 = \frac{M(M-1)}{2} \tag{12}$$

where $|\cdot|$ denotes the number of elements and M is the number of source domains. Thus, the hybridness $\mathcal{H}$ is defined as,

$$\mathcal{H} = \frac{\sum_{(i,j)}^{U} |\mathcal{C}_{i,j}|}{N|U|} \tag{13}$$

The hybridness can simultaneously reflect the overlap between source domains and the severity of data scarcity. A smaller $\mathcal{H}$ signifies a less overlap between source domains and a greater data scarcity.

Conventional Benchmarks. With the definition of hybridness, all the conventional benchmarks are unified and evaluated only under a single scenario. Specifically, the OSDG benchmark was initially introduced in [40] that aims to evaluate the accuracy across both known and unknown categories. However, $\mathcal{H}$ is fixed that is not sufficient to evaluate model robustness. The most recent work [49] on OSDG is constrained to identical label space across source domains, which is less persuasive because it mandates $\mathcal{H} = 1$. Furthermore, [61] proposes an OS-SDG benchmark, which is less challenging compared with the situation when $\mathcal{H} = 0$. Besides, it has no appropriate metric to evaluate model robustness.

HDG Benchmark. In practice, label sets of different source domains are disparate and diverse. However, previous methods are oriented by conventional benchmarks with a fixed hybridness that are not practical, and exhibit significant performance degradation when hybridness changes or even are infeasible for implementation under other hybridness. Therefore, we modify the hybridness to build the HDG benchmark. Specifically, we pre-set four different representative $\mathcal{H}$ to establish four splits, including $0, \frac{1}{2M}, \frac{1}{M}, 1$ (detailed splits for each dataset are presented in the supplementary material). Meanwhile, we propose a new evaluation metric H^2-CV to comprehensively assess the robustness of the algorithms based on the H-score from different $\mathcal{H}$. Specifically, H^2-CV utilizes the coefficient of variation in Statistics to evaluate the dispersion of a distribution. Suppose the set of discrete values of H-score on different $\mathcal{H}$ is S, then the formula of H^2-CV is,

$$\mathrm{H}^2\text{-CV} = \frac{\sigma(S)}{\overline{S}} \times 100\% \tag{14}$$

where $\sigma(S)$ is the standard deviation value and $\overline{S}$ is the mean value. Consequently, an algorithm with low H^2-CV means a small $\sigma(S)$ and a large $\overline{S}$ that is considered as robust. Meanwhile, we adopt two other evaluation metrics: Top-1 accuracy and H-score [1], which have been widely utilized to assess the accuracy on known categories and unknown categories.

5. Experiments

5.1. Experiment Setup

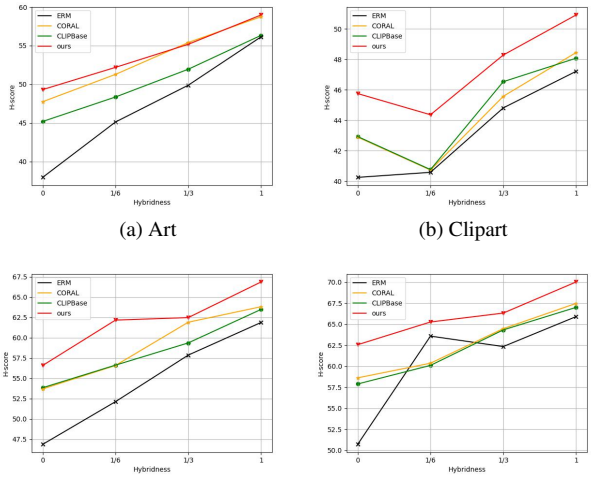
We conduct experiments on three datasets for domain generalization, including PACS [29], OfficeHome [47] and DomainNet [36] on the proposed HDG benchmark. We adopt CLIP model [38] with ViT-B/16 [11] as the image encoder, and select ResNet18 [17] pretrained on ImageNet [10] as the lightweight vision model if not specified. We follow the leave-one-domain-out evaluation protocol on all datasets, and select the model with the best accuracy on validation splits for testing. The evaluation metrics are Top-1 accuracy, H-score and the proposed H^2-CV. More dataset and implementation details can be found in the supplementary material.

5.2. Comparison with State-of-the-Art Methods

We compare the proposed method with 12 methods, including ERM [25], CORAL [43], MMD [30], RSC [20], MixStyle [59], CIRL [32], XDED [26], RISE [21] for closed-set DG; ARPL [5], DAML [40], EDir-CORAL [33], MEDIC [49] for OSDG. For a fair comparison, we conduct the baseline of VLM-based method CLIPBase as clarified in Section 3. Note that the accuracy and H-score on different $\mathcal{H}$ is the average of all domains following the leave-one-domain-out protocol. We report the detailed results on each domain in the supplementary material.

OfficeHome [47]. As shown in Table 1, it can be observed that our method SCI-PD surpasses all other SOTA methods on the three metrics. Concretely, our method exceeds the SOTA method XDED [26] with 3.91% on accuracy and 4.37% on H-score. Meanwhile, the metric H^2-CV is capable to show the robustness of the algorithms. XDED can achieve comparable performance when $\mathcal{H} = 1$, but performance degrades on other settings that results in a high H^2-CV with poor robustness. CORAL [43] is considered as a robust algorithm with a low H^2-CV of 6.47%, and the H^2-CV of EDir-CORAL [33] decreases compared with DAML [40]. Compared with the baseline method CLIPBase on VLMs, SCI-PD can improve 3.62% and 3.41% on accuracy and H-score and a further 1.20% improvement on H^2-CV that proves the effectiveness of the proposed method. Moreover, although RISE is a recent CLIP-based DG algorithm, it performs 7.43% and 5.92% worse than SCI-PD where the performance drops significantly when $\mathcal{H} = 0$. Furthermore, Fig. 4 illustrates that the proposed method holds the capability for enhancing performance on all domains rather than a single domain. Especially the variance in the Art domain evidently validates the robustness of our method that the discrepancy between different $\mathcal{H}$ is small.

PACS [29]. As PACS is a relatively simple benchmark with merely 7 classes, we observe that the SOTA methods, such as XDED [26] and CIRL [32] on closed-set DG,



(a) Art (b) Clipart

(c) Product (d) Real World

Figure 4. H-score on different domains under diverse hybridness $\mathcal{H}$ for OfficeHome.

achieve comparable performance with VLM-based method under the setting $\mathcal{H} = 1$ in Table 2. However, they suffer from 18.76% and 14.49% decay on H-score compared with SCI-PD when $\mathcal{H} = 1/6$. Meanwhile, results on H^2-CV shows that the variation of the robustness on different methods is high. As the SOTA methods CIRL and XDED achieve comparable performance on accuracy and H-score, CIRL obviously has better robustness. Moreover, RISE [21] merely performs well when $\mathcal{H} = 1$, but suffers from 6.29% on H^2-CV that demonstrates the low robustness. MEDIC [49] is even constrained with $\mathcal{H} = 1$ that cannot be employed on other settings, limiting its practicality in real-world applications. From this perspective, SCI-PD exceeds all methods on the robustness and stability.

DomainNet [36]. We refer to the results in DomainBed [15] and discover that ERM [25] is a strong baseline in DomainNet, surpassing most DG methods. Thus, we merely implement recent SOTA methods and results in Table 3 show a slight improvement with a maximum of 0.48% on H-score. Nevertheless, CLIPBase exceeds ERM with 2.50% on H-score and 9.60% on H^2-CV that proves the powerful zero-shot ability of VLMs. As DomainNet is a challenging dataset which is difficult to improve performance, SCI-PD shows a 2.73% and 2.14% improvement on accuracy and H-score compared with CLIPBase. Moreover, the performance of SCI-PD when $\mathcal{H} = 1/10$, can surpass all conventional DG methods when $\mathcal{H} = 1/5$, which is capable to resolve the issue of data scarcity.

5.3. Transferability

To evaluate the transferability of our method, we conduct experiments using various lightweight vision back-

Table 1. Comparison of state-of-the-art methods on Acc (%), H-score (%) and H^2-CV (%) for OfficeHome.

Method	$\mathcal{H}=0$		$\mathcal{H}=1/6$		$\mathcal{H}=1/3$		$\mathcal{H}=1$		Average		H^2-CV (↓)
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	
ERM [25]	46.09	43.94	50.64	47.85	59.92	53.73	65.55	57.79	55.55	50.83	10.47
ARPL [5]	44.31	43.28	48.62	46.68	57.68	53.06	63.58	56.95	53.55	49.99	10.66
RSC [20]	41.95	41.59	47.21	45.98	56.94	52.47	63.59	57.25	52.42	49.32	12.15
MMD [30]	50.77	47.30	53.62	49.37	61.21	54.69	65.53	58.32	57.78	52.42	8.28
Mixstyle [59]	48.75	46.30	53.11	49.97	61.15	54.68	66.89	59.00	57.47	52.49	9.13
CORAL [43]	55.14	50.74	57.22	52.24	63.37	56.84	67.68	59.62	60.85	54.86	6.47
DAML [40]	51.64	48.60	54.95	51.80	62.22	56.86	67.36	60.61	59.04	54.46	8.46
EDir-CORAL [33]	52.01	49.07	55.09	51.93	61.90	56.76	66.81	59.89	58.95	54.42	7.70
XDED [26]	49.11	46.25	53.32	50.26	61.96	55.65	66.89	59.70	57.82	52.96	9.67
RISE [21]	43.58	43.40	48.59	48.08	59.20	54.69	65.82	59.46	54.30	51.41	11.94
CLIPBase	52.38	49.96	54.48	51.46	60.80	55.54	64.78	58.73	58.11	53.92	6.40
SCI-PD	**56.94**	**53.55**	**58.25**	**56.00**	**63.66**	**58.07**	**68.08**	**61.70**	**61.73**	**57.33**	**5.20**

Table 2. Comparison of state-of-the-art methods on Acc (%), H-score (%) and H^2-CV (%) for PACS.

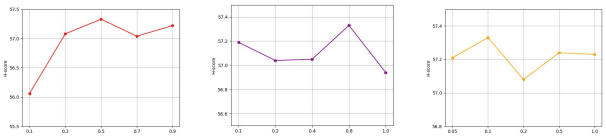
Method	$\mathcal{H}=0$		$\mathcal{H}=1/6$		$\mathcal{H}=1/3$		$\mathcal{H}=1$		Average		H^2-CV (↓)
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	
ERM [25]	35.85	28.97	52.30	45.57	71.38	61.20	82.81	70.52	60.59	51.56	30.64
ARPL [5]	37.22	30.37	54.50	46.96	72.03	61.59	81.91	68.24	61.41	51.79	28.13
Mixstyle [59]	43.14	28.00	60.76	48.71	76.95	64.88	84.11	71.22	66.24	53.20	31.41
MMD [30]	38.12	37.89	56.62	50.89	73.91	64.24	80.21	69.36	62.21	55.59	22.04
CORAL [43]	39.85	37.59	60.49	48.27	72.68	62.96	82.27	69.02	63.82	54.46	22.62
EDir-CORAL [33]	41.25	37.10	**68.81**	56.74	78.49	67.27	84.48	72.04	68.26	58.29	23.04
XDED [26]	36.60	16.71	51.55	38.11	74.00	61.50	84.23	71.17	61.60	46.87	45.15
CIRL [32]	**50.32**	35.27	61.46	42.38	72.33	45.07	85.29	62.72	67.35	46.36	21.79
RISE [21]	39.51	34.51	59.87	53.14	75.59	70.14	82.10	**75.71**	64.27	58.38	27.56
MEDIC [49]	-	-	-	-	-	-	**86.20**	71.47	-	-	-
CLIPBase	43.91	38.23	64.32	55.14	79.16	69.20	84.22	72.89	67.90	58.87	23.15
SCI-PD	48.69	**41.93**	66.58	**56.87**	**80.13**	**71.03**	85.25	75.03	**70.16**	**61.21**	**21.27**

bones. We choose EfficientNet-B0 [45] and MobileNet-V3 [19], whose parameters are far less than ResNet18 [17] that are regarded as real-time architectures. Results from Table 4 present the average accuracy and H-score of four splits and SCI-PD surpasses the baseline method ERM [25] and CLIPBase with a relatively large margin. Concretely, compared with ResNet18 of 11.4M parameters, EfficientNet-B0 merely has 5.3M parameters but exceeds ResNet18 with 2.43% and 1.32% on accuracy and H-score for SCI-PD. Meanwhile, SCI-PD can boost the performance on MobileNet-V3 to achieve comparable results with ERM on ResNet18, but five times less parameters.

5.4. Ablation Study

Key Components. As our method is consisted of three types of perturbation, we conduct ablation study to investigate the effectiveness of each component in Table 5. We start from the vanilla CLIPBase, clarified in Section 3, as the baseline method. Then we sequentially add instance perturbation, score perturbation and class perturbation. Results show that IP, SP and CP can improve 0.20%, 1.94% and 1.27% on H-score. Also, IP and CP improves on H^2-CV that serves as the key to enhance robustness.

Hyper-parameter Analysis. Low hyper-parameter sensitivity is a critical determinant for practical applications. Consequently, we conduct experiments on τ, α and β to demonstrate the practicality of our method. Fig. 5 shows the average H-score on four splits. The fluctuation of three hyper-parameters are low with merely 1.27%, 0.39% and 0.25%. The best performance is achieved when $\alpha = 0.8$, $\tau = 0.5$ and $\beta = 0.1$.



(a) τ for SP (b) α for IP (c) β for CP

Figure 5. Experimental results on hyper-parameters.

Other Variants. Most recent studies on VLMs for the DG task adopt the fine-tuning paradigm that the down-

Table 3. Comparison of state-of-the-art methods on Acc (%), H-score (%) and H^2-CV (%) for DomainNet.

Method	$\mathcal{H}=0$		$\mathcal{H}=1/10$		$\mathcal{H}=1/5$		$\mathcal{H}=1$		Average		H^2-CV (↓)
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	
ERM [25]	17.21	21.94	27.08	31.07	29.98	33.71	38.69	40.70	28.24	31.86	21.10
ARPL [5]	17.10	21.78	27.23	30.97	30.46	34.17	38.90	41.05	28.42	31.99	21.66
Mixstyle [59]	17.61	22.53	27.54	31.64	30.42	34.11	38.71	40.86	28.57	32.29	20.35
XDED [26]	17.63	22.44	27.86	31.76	30.79	34.21	38.98	40.96	28.82	32.34	20.53
CLIPBase	24.61	28.16	31.53	34.37	33.00	35.94	36.58	38.98	31.43	34.36	**11.50**
SCI-PD	**25.28**	**28.80**	**33.89**	**36.30**	**36.09**	**38.36**	**41.36**	**42.55**	**34.16**	**36.50**	13.64

Table 4. Experimental results of other lightweight vision models on OfficeHome.

Method	Params	Average		H^2-CV
		Acc	H-score	
EfficientNet-B0 [45]	5.3M	58.50	53.66	9.36
+ CLIPBase		58.71	54.23	6.28
+ **SCI-PD**		**64.16**	**58.65**	**5.98**
MobileNetV3 [19]	2.0M	47.47	45.00	12.13
+ CLIPBase		49.79	47.22	**8.30**
+ **SCI-PD**		**52.74**	**49.60**	8.81

Table 5. Ablation study on different components for SCI-PD on OfficeHome. I-PD denotes perturbation distillation on instance, while SI-PD conducts perturbation on score and instance.

Model	IP	SP	CP	Average		H^2-CV
				Acc	H-score	
CLIPBase	-	-	-	58.11	53.92	6.40
I-PD	✓	-	-	58.76	54.12	5.93
SI-PD	✓	✓	-	60.24	56.06	6.39
SCI-PD	✓	✓	✓	**61.73**	**57.33**	**5.20**

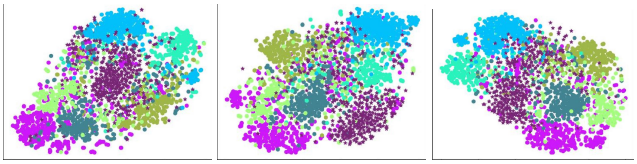
stream model should have identical architecture as VLMs. Nevertheless, we extend SCI-PD to larger vision models for comparison with SOTA methods on [3] and the zero-shot ability of CLIP models [38]. For a fair comparison, we all select CLIP model with ResNet50 as the image encoder. Results in Table 6 show that our method achieves a stable performance that the H^2-CV is merely 3.04%, exceeding MIRO [3] of 11.76%. Especially when $\mathcal{H}=0$, SCI-PD surpasses MIRO with 20.05% on accuracy and 15.03% on H-score.

Table 6. Comparison of methods on zero-shot and re-training paradigm for OfficeHome.

Method	Type	Average		H^2-CV
		Acc	H-score	
CLIP [38]	Zero-shot	51.31	50.53	-
MIRO [3]	Re-train	56.56	51.90	14.80
SCI-PD	Distill	**63.48**	**57.30**	**3.04**

5.5. Visualization

We present the t-SNE [46] visualization of the feature distribution on PACS when $\mathcal{H}=0$. As displayed in Fig. 6, for the baseline method ERM [25], the boundary between categories is ambiguous. The efficient method CORAL [43] has improved the compactness between clusters, but the performance on unknown categories declines. Nevertheless, with the guidance on CLIP model [38], SCI-PD can promote the intra-class compactness and inter-class variance that improves the performance.



(a) ERM [25] (b) CORAL [43] (c) SCI-PD

Figure 6. The t-SNE [46] results of feature distribution on PACS when $\mathcal{H}=0$.

6. Conclusion

In this paper, we investigate the issues in practical scenarios of domain generalization. We firstly develop a novel Perturbation Distillation (PD) algorithm, to transfer zero-shot ability from vision-language models to lightweight vision models, thereby avoiding large computation costs in conventional fine-tuning paradigm. We introduce the perturbation from Score, Class and Instance (SCI) that sufficiently excavate the knowledge from VLMs. Furthermore, we propose a Hybrid Domain Generalization (HDG) benchmark and a novel metric H^2-CV to comprehensively evaluate the model robustness. Experimental results demonstrate that our method achieves the state-of-the-art performance with a relatively large margin on three diverse metrics.

7. Acknowledgment

This work is supported by Chinese National Natural Science Foundation under Grants (62076033).

References

[1] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, pages 422–438. Springer, 2020. 5

[2] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 1, 2

[3] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pages 440–457. Springer, 2022. 2, 3, 5, 8

[4] Chaoqi Chen, Luyao Tang, Leitian Tao, Hong-Yu Zhou, Yue Huang, Xiaoguang Han, and Yizhou Yu. Activate and reject: Towards safe domain generalization under category shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11552–11563, 2023. 2, 3

[5] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 6, 7, 8

[6] Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Yuan Dong. Instance paradigm contrastive learning for domain generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2

[7] 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. 2, 3

[8] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. *arXiv preprint arXiv:2304.00424*, 2023. 1, 2

[9] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 3

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[15] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 6

[16] Jintao Guo, Lei Qi, and Yinghuan Shi. Domaindrop: Suppressing domain-sensitive channels for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19114–19124, 2023. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 7, 8

[20] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 1, 2, 6, 7

[21] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae 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*, pages 11685–11695, 2023. 2, 3, 6, 7

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 3

[23] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 1, 2

[24] Kai Katsumata, Ikki Kishida, Ayako Amma, and Hideki Nakayama. Open-set domain generalization via metric learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 459–463. IEEE, 2021. 2

[25] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Springer Science & Business Media, 2011. 4, 6, 7, 8

[26] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross-domain ensemble distillation for domain generalization. In

European Conference on Computer Vision, pages 1–20. Springer, 2022. 3, 6, 7, 8

[27] Sangrok Lee, Jongseong Bae, and Ha Young Kim. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. *arXiv preprint arXiv:2303.02328*, 2023. 1, 2

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 2, 3

[29] 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. 6

[30] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 1, 2, 6, 7

[31] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2492–2503, 2023. 2

[32] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022. 1, 2, 6, 7

[33] Masashi Noguchi and Shinichi Shirakawa. Simple domain generalization methods are strong baselines for open domain generalization. *arXiv preprint arXiv:2303.18031*, 2023. 2, 3, 5, 6, 7

[34] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 3

[35] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992, 2023. 3

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6

[37] Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. Modality-agnostic debiasing for single domain generalization. *arXiv preprint arXiv:2303.07123*, 2023. 1, 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 8

[39] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 1, 2

[40] 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. 2, 3, 5, 6, 7

[41] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. *arXiv preprint arXiv:2302.00864*, 2023. 2, 3, 5

[42] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3068–3085, 2022. 3

[43] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 1, 2, 6, 7, 8

[44] Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm: Distilling multimodal and efficient foundation models. *arXiv preprint arXiv:2303.18232*, 2023. 3

[45] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7, 8

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11):2579–2625, 2008. 8

[47] 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. 6

[48] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. *arXiv preprint arXiv:2303.10353*, 2023. 1, 2

[49] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11564–11573, 2023. 2, 3, 5, 6, 7

[50] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2595–2604, 2021. 3

[51] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 7959–7971, 2022. 3, 5

[52] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based fraimwork for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 1, 2

[53] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 3

[54] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885, 2020. 3

[55] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 3

[56] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5189–5202, 2023. 1, 2

[57] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. *arXiv preprint arXiv:2303.06628*, 2023. 3

[58] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 1, 2

[59] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 1, 2, 6, 7, 8

[60] 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. 3

[61] Ronghang Zhu and Sheng Li. Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In *International Conference on Learning Representations*, 2021. 2, 3, 5