A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance

Zeyi Huang[1] Andy Zhou[2] Zijian Lin[3] Mu Cai[1] Haohan Wang[2†] Yong Jae Lee[1†]
[1]University of Wisconsin-Madison [2]University of Illinois Urbana-Champaign [3]Imperial College London
{zeyihuang, mucai, yongjaelee}@cs.wisc.edu {andyz3,haohanw}@illinois.edu z.ling22@imperial.ac.uk

Abstract

Domain generalization studies the problem of training a model with samples from several domains (or distributions) and then testing the model with samples from a new, unseen domain. In this paper, we propose a novel approach for domain generalization that leverages recent advances in large vision-language models, specifically a CLIP teacher model, to train a smaller model that generalizes to unseen domains. The key technical contribution is a new type of regularization that requires the student's learned image representations to be close to the teacher's learned text representations obtained from encoding the corresponding text descriptions of images. We introduce two designs of the loss function, absolute and relative distance, which provide specific guidance on how the training process of the student model should be regularized. We evaluate our proposed method, dubbed RISE (Regularized Invariance with Semantic Embeddings), on various benchmark datasets, and show that it outperforms several state-of-the-art domain generalization methods. To our knowledge, our work is the first to leverage knowledge distillation using a large vision-language model for domain generalization. By incorporating text-based information, RISE improves the generalization capability of machine learning models.

1. Introduction

An image is worth a thousand words, indeed, because of its power to convey a wealth of information through its visual details. However, a well-written sentence, on the other hand, has the power to concisely capture the essential information that is common to many different images. By describing a scene with a few carefully chosen words, a writer can create a mental image in the reader's mind that conveys the essence of what is being depicted. This perspective is particularly useful when communicating information efficiently, or when emphasizing a specific scene aspect without getting bogged down in extraneous details. Thus, we suggest that a sentence speaks a thousand images.

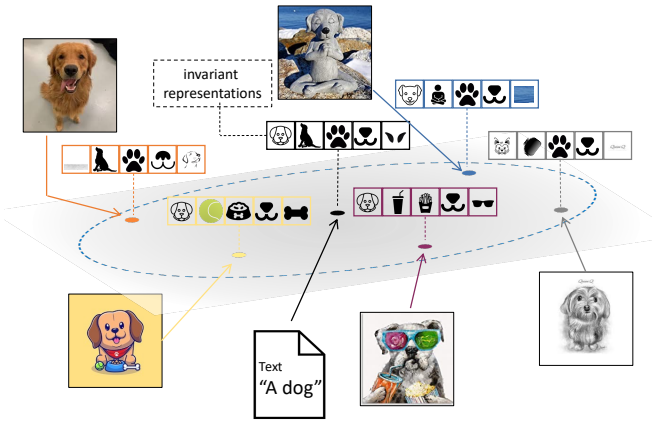

Figure 1. The key intuition behind our argument. While images can capture more details, text can directly summarize the core concept to represent the object of interest.

Essential semantic information delivered by an image plays a pivotal role in helping models generalize to shifted distributions, whereas other detailed information (e.g., in the background not relevant to the main object) captured in images may not be as effective for this purpose. The study of domain generalization [37] investigates the problem of training a model with samples from several domains (or distributions) and then testing the model with samples from a new, unseen domain. The training domains are commonly referred to as source domains, and the test domain is referred to as the target domain. Previous studies have identified a challenge in training effective domain generalization models due to the models' tendency to learn domain-specific features [11]. Consequently, numerous works have focused on regularizing the models to learn representations that are invariant to domain-specific features [29, 31, 80, 56, 36, 4, 1, 10, 39, 47, 17]. This regularization ensures that the models extract features that are

† equal advising. Code is available at github.com/OoDBag/RISE

common to multiple domains and are therefore more likely to generalize to unseen domains. By mitigating the influence of domain-specific features, the idea is to improve the generalization capability of these models and ensure that they perform well on a variety of different domains.

In this paper, we build upon this line of research by investigating methods for learning domain-invariant features in machine learning models. Our proposed method is inspired by a simple intuition: while an image tends to convey rich but sometimes excessive details through its pixels, a corresponding text description can describe the crux of the image content in a highly concise and complementary manner; see Figure 1. Therefore, the most effective regularization might involve incorporating a regularization strategy in which the learned representations need to be close to the representations obtained from encoding the corresponding concise text descriptions of an image.

Building upon this argument, we propose a novel domain generalization approach that leverages recent advances in vision-language models, such as CLIP [46], to train our domain generalization models. We are particularly interested in the setting where our final models are relatively small, and thus, can benefit from a large pre-trained vision-language teacher model through distillation. Our method, dubbed RISE (Regularized Invariance with Semantic Embeddings), incorporates both the vision and language components of a pre-trained and frozen CLIP teacher, inspired by the importance of the representations encoded by the language component. Specifically, RISE includes three loss functions: the empiricial risk minimization (ERM) loss that follows the standard pipeline of domain generalization, the model distillation loss that leverages the pretrained weights of the image component of CLIP, and the cross-domain (text to image) distance loss that uses the power of text through the language component of CLIP.

To fully harness the power of language, we introduce two different designs of the cross-domain distance loss function: the absolute distance design pushes the student's learned representation closer to the teacher's domain-invariant representation learned from language, while the relative distance design enforces that the relative domain distances in the teacher's encoded language space are transferred over to the learned representation in the student's encoded image space.

Contributions. In summary, our main contributions are:

- To the best of our knowledge, we are the first to leverage knowledge distillation using a large vision-language model as a teacher for domain generalization.
- We propose to regularize the representation learned by the student through images to be closer to the ones from the teacher's text representation, as text can be more concise and capture the semantic essence.
- We propose two loss functions, namely the absolute distance and the relative distance, which provide specific guidance on how the student model's training process should be regularized.
- We conduct a rich set of experiments to validate the effectiveness of our model RISE on domain generalization benchmarks and ablate the performance of each of its components.

2. Related Work

2.1. Domain Generalization

Domain Generalization [37] has been widely studied in recent years. It mainly studies the problem of training a model from the data collected from multiple source distributions and then testing the trained model on a target distribution that is different from the training ones. Because of this problem formulation, a natural assumption to guide the development of the methods is that if the model can learn a representation that is invariant across the multiple training domains, it will generalize well to the unseen test domain. A large number of methods have been invented following this natural assumption, aiming to force the invariance across samples of training distributions, either through explicit regularization based methods [29, 31, 80, 56, 36, 4, 1, 10, 39, 47, 17, 83, 9, 58, 62, 35, 27, 77, 69] or (virtual) data augmentation methods [49, 75, 12, 82, 21, 61].

In addition, the assumption above of "invariance across multiple domains" is being challenged in recent years with the argument that a more realistic scenario is when the training datasets are not necessarily partitioned into multiple distributions/domains with clear boundaries during training. As a response to this argument, more powerful methods that do not rely on the domain partitions to force invariance have been introduced [23, 53, 22]. Our method in this paper in tested in the context of this challenging scenario.

Also, in recent years, it seems the community is using the terminology *out-of-distribution (OOD) generalization* to largely refer to domain generalization. For more detailed discussions on the topics of domain generalization and out-of-distribution (OOD) generalization, we refer the reader to dedicated surveys [57, 51].

More closely related to our contribution in this paper, we notice a prior work that also leverages a pre-trained model to improve domain generalization performance. Specifically, Domain Prompt Learning (DPL) [79] utilizes a lightweight prompt adaptor to automatically generate a prompt that estimates domain-specific features given unlabeled examples from each distribution. The following two works are not closely related to CLIP but leverage CLIP as their pre-trained model for domain generalization: [32] dispatches proper pre-trained models (including CLIP) to each sample based on their generalization ability. [6] re-

formulates the DG objective by mutual information with oracle models (including CLIP).

Key novelty: Unlike prior work, we leverage CLIP as a teacher and regularize the student's learned representation through images to be closer to the corresponding text representation of the teacher. Our method includes two loss functions that directly leverage language for learning invariant image representations.

2.2. Knowledge Distillation

Knowledge distillation is a technique for transferring knowledge from a teacher model to a student model, by optimizing the student model to match the outputs or intermediate features of the teacher model. This technique is used in numerous distinct contexts, such as semi-supervised learning [40] or even self-supervised learning [63].

Ever since the introduction of the term in [20], a plethora of techniques have been developed [13] with improvement in various aspects, centering around the idea of how to align the output of the student model to the teacher model for every input, where the alignment and output are both subject to various concrete definitions. For example, one branch of works varies on how to enforce the alignment, with a particular focus on the loss function design over the outputs between the teacher and the student for every sample, with popular studies such as ℓ_1 [26], ℓ_2 [7, 44, 59], MMD [24], KL divergence [8, 43, 42], and cross-entropy losses [64, 33]. Another branch studies how to define the output, which, at a high-level, has variants of directly using the embeddings from a certain (or final) layer [67, 15, 18, 50], or some more structured functions of the (pair-wise) embeddings of those layers [30, 76, 72]. There are also other branches such as the student-teacher architecture design or distillation algorithms that are not directly related to our study in this paper; we recommend the reader to refer to a survey for more details [13].

Among these works, the most relevant technical development to our method is to distill the preservation of the relationship between samples from the teacher model to the student model. For example, [7] distills while the the student also learns the relationship between samples after the relationship is projected to a lower dimensional space, and other works more directly optimize the similarity of the pair-wise distances between embeddings after each pair of samples is fed into the teacher and student models, respectively [74, 45, 34, 41].

Key novelty: The objective of prior work KDDG [60] is to distill the knowledge of a *pure vision teacher* to a student model. In contrast, our approach focuses on distilling the knowledge of large-scale *vision and language* models (CLIP) to the student model.

2.3. Large Vision-Language Models

Recent advances in vision-language models [46, 25, 66, 68, 65, 73] have shown promising results in learning generic visual representations and facilitating zero-shot transfer to diverse downstream classification tasks through the use of prompts. These models typically rely on a contrastive loss to align a visual encoder and a text encoder in a shared feature space. Trained on large-scale image-text pairs, these vision-language models demonstrate transferability across a wide range of applications, including object detection [14], semantic segmentation [81], and point cloud classification [78].

In particular, Contrastive Language Image Pre-training *i.e.*, CLIP [46] utilizes 400M pretraining image-text pairs to conduct image-caption contrastive pretraining. Empirically, CLIP shows superior zero-shot image classification performance, achieving 76.2% top-1 accuracy on the ImageNet validation set, which is on par with the performance of an ImageNet fully-supervised ResNet101 model. Furthermore, CLIP shows potential domain generalization capabilities. For example, it achieves 60.2% accuracy on ImageNet Sketch Dataset while the ImageNet supervised training model (ResNet101) can only achieve 25.2% accuracy. This motivates us to answer the following question: *What is the best way to distill CLIP's rich domain knowledge to a smaller student network for domain generalization tasks?*

3. RISE: Regularized Invariance with Semantic Embeddings

In this section, we present the details of our approach for distilling a large vision-language model's learned semantic knowledge into a smaller student model for domain generalization. Importantly, we use a *pre-trained and frozen* CLIP [46] as the teacher in this work.

3.1. Notations, Baseline, and Distillation from Teacher's Image Component

We first introduce our notations. We use $(\mathbf{X}, \mathbf{Y})$ to denote the training dataset with n (data,label) paired samples. these data samples can be from multiple domains or distributions, but since our model does not need the domain-ID information, we do not need a notation to specify which distribution the samples are from. Let $(\mathbf{x}, \mathbf{y})$ denote one sample and $f(\cdot;\theta)$ denote the model we aim to train. Thus, a vanilla paradigm of training a domain generalization model without domain IDs is as simple as the standard empirical risk minimization (ERM):

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} l(f(\mathbf{x};\theta),\mathbf{y}), \tag{1}$$

where $l(\cdot,\cdot)$ denotes a generic loss function.

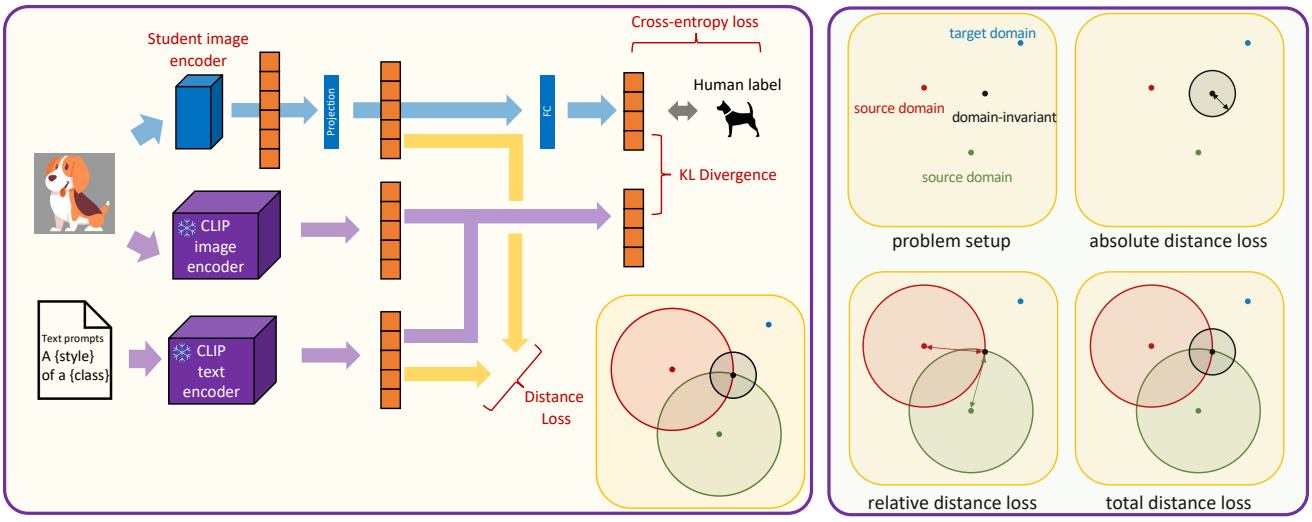

Figure 2. **(Left)** Overview of the pipeline of our proposed method as a combination of three losses: the cross-entropy loss as in standard supervised training, the KL divergence loss as in domain distillation, and our proposed cross-domain (text to image) distance loss. Here, a *pre-trained and frozen* CLIP is the teacher model. The teacher model is not trained. **(Right)** The intuition of our two proposed losses and their combined effects. (i) In latent space, we aim to regularize the model to learn a representation that is close to the domain-invariant representation from the teacher's text space; (ii) the absolute distance loss can regularize the search to be within the shaded area; (iii) the relative distance loss can regularize the search to be within the overlap area; (iv) the combined loss can shrink the search space by overlapping these two.

We aim to incorporate rich prior knowledge from a CLIP pretrained image model teacher through distillation. We use $h(\cdot;\phi)$ to denote this pretrained model, and the training process as:

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} l(f_l(\mathbf{x};\theta), h_l(\mathbf{x};\phi)), \tag{2}$$

where $l(\cdot,\cdot)$ denotes a typical loss function that measures the distance of two vectors (e.g., KL divergence between the two predicted output distributions). $f_l(\cdot;\theta)$ and $h_l(\cdot;\phi)$ denote the output of logits instead of the final prediction.

To use CLIP to produce output distributions in a classification setting, we feed in the text encoder of CLIP with queries generated with the label of the images (one query per label) such as "a photo of a dog", "a photo of an elephant", etc. We use the image encoder's embedding to match each of the class query embeddings of the text encoder using cosine similiarity, and normalize the result to generate the output logits.

3.2. Regularization with Teacher's Language Component

We use $g_l(\cdot;\delta)$ to denote the CLIP teacher's language component that takes the input text phrase and generates an embedding. In general, if we have a generic description of the image, such as "$\mathbf{z}$ = a photo of a dog", we can directly feed this text phrase into the model to generate the corresponding embedding, following $\mathbf{e_z}(\text{dog}) = g_l(\mathbf{z};\delta)$.

However, in practice, although "a photo of a dog" is recommended by CLIP as a standard text template, this text might not be generic enough as it still indicates the pixel statistics of the image following the typical statistics of what a *photo* has, which is potentially different from what a *sketch* or what a *painting* has.

To overcome the potential over-reliance on the pixel statistics of *photo*, we use the recommended list of eighty templates of text prompts by CLIP [46], including from "a photo of my {}", "an art of {}", to even "a tattoo of the {}" and consider their averaged representation as the generic teacher's text representation of the depicted object.

More concretely, we build the generic representation of class i by

$$\mathbf{e_z}(i) = \frac{1}{n} \sum_{\mathbf{z}\in\mathbf{Z}(i)} g_l(\mathbf{z};\delta)$$

where $\mathbf{Z}(i)$ denotes the set of recommended text templates when the class is filled in with the class name corresponding to object i.

With the teacher's generic text embedding $\mathbf{e_z}(i)$, we aim to regularize the learning process of the student model to match its learned image representation to this generic representation, with two losses that function differently: absolute distance loss and relative distance loss.

3.2.1 Absolute Distance Loss

The absolute distance loss is designed to directly push the student's learned image representation to be close to the teacher's generic text representation:

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \sum_{i\in\mathbf{I}} \mathbb{I}[c(\mathbf{x})=i] k(f_l(\mathbf{x};\theta), \mathbf{e}_\mathbf{z}(i)), \quad (3)$$

where $k(\cdot,\cdot)$ is a distance metric, $\mathbf{I}$ is the collection of all possible classes, and $\mathbb{I}[c(\mathbf{x})=i]$ is simply an identity function that returns 1 if the class of $\mathbf{x}$ is i and 0 otherwise.

Ideally, if we can achieve the minimum value from (3), we will train a student model that can learn generic visual representations that are likely to be invariant across the input domains and perform well on target domains.

However, in practice, due to the difficulties of optimizing deep learning models on real-world data, the optimization cannot easily find such optimal solutions. Therefore, we need to introduce an additional regularization to help shrink the search space.

3.2.2 Relative Distance Loss

We introduce a relative distance loss that can better describe where the target generic representation is.

To do so, we need to first introduce several additional anchor points. For a domain generalization problem with possible training domain $d \in \mathbf{D}$, and for every class $i \in \mathbf{I}$, we generate $\mathbf{e}_\mathbf{z}(d,i)$ by feeding the text prompt "a $\{$d$\}$ of $\{i\}$" to the teacher's text encoder.

Therefore, we have the relative distance loss as

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \sum_{i\in\mathbf{I}} \mathbb{I}[c(\mathbf{x})=i] \sum_{d\in\mathbf{D}} k_1\Big(k_2\big(f_l(\mathbf{x};\theta), \mathbf{e}_\mathbf{z}(d,i)\big), k_2\big(\mathbf{e}_\mathbf{z}(i), \mathbf{e}_\mathbf{z}(d,i)\big)\Big), \quad (4)$$

where k_1 and k_2 denote two distance metrics.

Intuitively, the relative distance loss helps to pinpoint the location of the teacher's generic text representation by pushing the relative position of the student's learned representation from images with respect to those anchor points to be the same as the position of the generic representation with respect to the anchor points.

The idea of the relative distance loss is to help the model to get to the generic embedding more directly. How it can help in searching for the generic representation is illustrated in the right-hand side of Figure 2.

3.3. Full Method

Connecting all the pieces above, our full method is to train the model with the loss functions from (1) to (4), with hyperparameters to balance the contribution of each method; see the left-hand side of Figure 2.

Formally, our final method is to train the model with

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} \lambda_1 l(f(\mathbf{x};\theta),\mathbf{y}) + \lambda_2 l'(f_l(\mathbf{x};\theta), h_l(\mathbf{x};\phi)) + \sum_{i\in\mathbf{I}} \mathbb{I}[c(\mathbf{x})=i]\lambda_3 \Bigg(k(f_l(\mathbf{x};\theta), \mathbf{e}_\mathbf{z}(i)) + \sum_{d\in\mathbf{D}} k_1\Big(k_2\big(f_l(\mathbf{x};\theta), \mathbf{e}_\mathbf{z}(d,i)\big), k_2\big(\mathbf{e}_\mathbf{z}(i), \mathbf{e}_\mathbf{z}(d,i)\big)\Big)\Bigg)$$

where λ_1, λ_2, and λ_3 are three hyperparameters that balance each loss term.

3.4. Implementation Details

In practice, we implement l as cross-entropy loss, l' as KL divergence, k as CosineSimilarity, k_1 as CosineSimilarity, and k_2 as L2 loss. The KL divergence l' introduces one more hyperparameter temperature t to control the smoothness of CLIP's predictions. The detail of distance metrics selection is analyzed in the ablation study. We use one linear layer to project the image embedding of the student to the text embedding of the CLIP teacher. During inference, images are passed through the student image encoder and FC layer (or CLIP's text embedding) to make final predictions.

4. Experiments

We evaluate our approach in leveraging language as a regularization strategy for training a student image model for domain generalization. We compare state-of-the-art domain generalization methods and perform ablation studies to analyze the various components of our model.

4.1. Setup

We follow the setting in [16, 71] and evaluate our domain generalization approach. Specifically, we use the same strategy for model selection, dataset splitting, and network backbone.

4.2. Datasets, Hyperparameter Search, and Model Selection

We follow DomainBed [16] and Ood-bench [71] to choose datasets that cover as much variety as possible from the various OOD research areas for our experiments. We conduct experiments on four OOD datasets: Terra Incognita [3], OfficeHome [55], VLCS [54], and PACS [28].

To be consistent with the existing line of work, we use the training-validation protocol for model selection: given N domains, it uses 90% the amount of data in $N-1$ domains for training, the other 10% for validation, selects the best model based on the validation result, tests the model on the held-out domain and reports this result.

Table 1. Results of domain generalization methods with ResNet backbone. Ens/MA stands for Ensemble/ Moving Average. $\star$ denotes fine-tuning on target datasets. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss. MT stands for Mix Teacher engineering technique. We report averaged accuracy across three runs.

Method	Backbone	Ens/MA	PACS	VLCS	OfficeHome	Terra	Ave
ERM [70]	ResNet18	No	81.5	73.2	63.3	43.6	65.4
Best SoTA competitor	ResNet18	No	83.4 [22]	74.1 [22]	63.8 [29]	44.5 [22]	66.5
ERM [16]	ResNet50	No	85.7	77.4	67.5	47.2	69.5
Best SOTA competitor	ResNet50	No	86.6 [48]	78.8 [52]	68.7 [52]	48.6 [38]	70.7
Ensemble [2]	ResNet50	Yes	87.6	78.5	70.8	49.2	71.5
SWAD [5]	ResNet50	Yes	88.1	79.1	70.6	50.0	71.9
EoA [2]	ResNet50	Yes	88.6	79.1	72.5	52.3	73.1
CLIP [79] (Teacher)	ViT B/16	No	96.1	82.3	82.3	50.2$\star$	77.7
ERM + Hint	ResNet18	No	84.6	78.0	64.6	47.0	68.6
ERM + Hint + AD	ResNet18	No	85.1	78.5	65.6	48.2	69.4
ERM + Hint + RD	ResNet18	No	84.9	78.2	65.2	47.9	69.0
ERM + Hint + AD + RD (Our full method)	ResNet18	No	85.3	78.6	65.9	48.4	69.6
ERM + Hint	ResNet50	No	88.4	80.7	70.2	50.5	72.5
ERM + Hint + AD	ResNet50	No	89.0	81.5	71.3	52.2	73.5
ERM + Hint + RD	ResNet50	No	88.8	81.2	71.1	51.7	73.2
ERM + Hint + AD + RD (Our full method)	ResNet50	No	89.4	81.7	71.6	52.3	73.8
ERM + Hint + AD + RD + MT (Our full method)	ResNet50	Yes	90.2	82.4	72.6	54.0	74.8

There are altogether four hyperparameters for our method – the weights of supervised loss λ_1, distillation loss λ_2, distance losses λ_3, and temperature t. Overall, we set the hyperparameter search space of our method as $\lambda_1 \in [0.1, 1.0]$, $\lambda_2 \in [0.1, 1.0]$, $\lambda_3 \in [0.1, 1.0]$, $t \in [1.0, 3.0]$. We adopted the same hyperparameter search protocol used in [16, 70].

4.3. Empirical Results

Zero-shot performance of CLIP teacher. We select CLIP ViT-B/16 as the teacher for the following experiments (due to limited computational resources we could not try larger models). In Table 1, CLIP ViT-B/16 achieves 96.1%, 82.3%, 82.3%, and 34.1% on PACS, VLCS, Office-Home, and Terra Incognita respectively when performing zero-shot inference. Except for Terra Incognita, both CLIP models outperform the best individual state-of-the-art results by up to 7%. Because of the extremely low zero-shot accuracy on Terra, we finetune CLIP on Terra to obtain a better CLIP teacher, which achieves 50.2% zero-shot accuracy. Overall, we use finetuned CLIP ViT B/16 for Terra Incognita and zero-shot CLIP ViT B/16 for the remaining three datasets.

Comparison with existing DG methods We compare to the recent top DG algorithms, using both ResNet18 and ResNet50 pre-trained on ImageNet-1k as the backbone. Our results are presented in Table 1. The "Best SoTA competitor" refers to the highest performance in the literature within the standard DG experimental protocol, and the numbers listed under this category may be from different methods. In addition, we also include ensemble and weight-averaging techniques in the third-row panel.

We first study the effect of the standard distillation loss [19]. We use the soft labels produced by the CLIP teacher as an additional target for the student to match, in addition to the (one-hot) ground-truth labels. This is done by minimizing the KL-divergence between the predicted distributions of the student and the teacher. Training a student with human labels and distillation loss (ERM + Hint in the last two panels), already outperforms most of the state-of-the-art methods on the benchmarks. EoA [2], a moving average variant method, is the only method that outperforms ERM + Hint with the ResNet50 backbone. Next, we study the effect of our proposed method. We observe that adding absolute distance (AD) and relative distance (RD) to ERM + Hint both result in clear performance gains, and together produce the best results which indicate their complementarity. For ResNet 18, AD, RD, and AD + RD provide 0.8%, 0.4%, and 1.0% improvement over ERM + Hint respectively. For ResNet 50, AD, RD, and AD + RD provide 1.0%, 0.7%, and 1.3% improvement over ERM + Hint respectively.

4.4. Ablation Studies

In this section, we study the impact of each component in our method. We evaluate our method with a ResNet50 backbone on the most popular DG benchmark PACS to conduct the following experimental analyses.

4.4.1 Impact of using text embedding as supervision

We study the impact of using CLIP's text embedding and image embedding as supervision for our absolute distance loss and relative distance loss. The results displayed in Table 2 indicate that for both our absolute distance loss and relative distance loss, utilizing text embedding of CLIP as supervision yields better results compared to using the image embedding counterpart for regulating the learning pro-

Method	Embedding	Acc
ERM + Hint + AD	Image	88.4
ERM + Hint + AD	Text	89.0
ERM + Hint + RD	Image	88.1
ERM + Hint + RD	Text	88.8

Table 2. Analysis of using CLIP's image embedding and text embedding as supervision. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss.

cess of our model, despite having the same loss function setting. Specifically, ERM + Hint + AD and ERM + Hint + RD with text embedding supervision outperform their image embedding counterparts with 0.6% and 0.7% improvement respectively. This analysis helps validate our assumption that CLIP's text embedding contains rich semantic information, and it can be treated as a domain-invariant representation since it is independent of images. In addition, Table 2 demonstrates that both our absolute distance loss and relative distance loss exhibit comparable performance under the ERM + Hint setting. Specifically, ERM + Hint achieved 88.4% (+AD) and 88.1% (+RD) using image embedding. Under ERM + Hint setting with text embedding, absolute distance loss performs slightly better than relative distance loss where ERM + Hint attains 89.0% (+AD) and 88.8% (+RD).

4.4.2 Impact of each loss component

Method	Acc
ERM	85.7
ERM + Hint	88.4
ERM + AD	87.8
ERM + RD	87.2
ERM + Hint + AD	89.0
ERM + Hint + RD	88.8
ERM + Hint + AD + RD	89.4

Table 3. Analysis of the effectiveness of each loss function in our method using ResNet50 backbone on PACS. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss.

Table 3 demonstrates that each component of our loss function contributes to the final performance. By adding one additional loss component to ERM (85.7%), ERM + Hint (88.4%), ERM + absolute (87.8%), and ERM + relative (87.2%), all get substantial improvements: +2.7%, 2.1%, 1.5%, respectively. Interestingly, ERM + AD achieves comparable performance with ERM + Hint which suggests that using CLIP's text embedding as supervision has the potential to match the performance of using the entire CLIP model. That is, a CLIP teacher can be used to generate supervisory signals for distillation without having access to any images. Moreover, by adding absolute distance loss and relative distance loss to ERM + Hint, there are further improvements of 0.6% and 0.4%, respectively, for ERM + Hint + AD and ERM + Hint + RD. Finally, by combining all components and using ERM + Hint + AD + RD (89.4%), we observe a significant improvement of 3.7% compared to using ERM only (85.7%).

4.4.3 Impact of prompt engineering and ensemble

Method	Template	Acc
ERM + Hint + AD	a photo of a $\{class\}$	88.5
ERM + Hint + AD	Ensemble template	89.0
ERM + Hint + RD	a photo of a $\{class\}$	88.3
ERM + Hint + RD	Ensemble template	88.8

Table 4. Analysis of CLIP's prompt engineering end ensemble. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss.

Table 4 demonstrates the effectiveness of having a prompt ensemble template, which enhances the accuracy compared to a single prompt template. Both ERM + Hint + AD and ERM + Hint + RD settings display an accuracy improvement of 0.5%. The ensemble template utilizes 80 representative templates of text prompts by CLIP [46]. The improvement in accuracy suggests that the text embedding generated by the ensemble template is more robust than the single template counterpart (i.e., "a photo of a {}") when facing distribution shift tasks. For those interested in exploring the details of the eighty prompt ensemble templates, they can be found here.

4.4.4 Impact of different distance metrics

Method	Loss	Acc
ERM + Hint + AD	CosineSimilarity	89.0
ERM + Hint + AD	Supervised Contrastive	88.6
ERM + Hint + AD	L1	88.0
ERM + Hint + AD	L2	88.1
ERM + Hint + RD	KL	88.7
ERM + Hint + RD	L1	88.3
ERM + Hint + RD	L2	88.8

Table 5. Effect of different regularization and distance metrics. For ERM + Hint + RD, we fix k_2 to be consine similarity, and only explore which kind of distance metric k_1 works the best with k_2. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss.

Table 5 shows the variation in performance due to different regularization and distance metrics. When considering the ERM + Hint + absolute distance loss setting (i.e., choosing the distance metric k in Eqn. 3), the Cosine Similarity loss (89.0%) outperforms the Supervised Contrastive

(88.6%), L1 (88.0%), and L2 (88.1%) approaches. (For Supervised Contrastive, the positive/negative pairs are the student's image feature and teacher's text feature for the ground-truth/non-ground-truth class.) On the other hand, for the ERM + Hint + relative distance loss setting (i.e., choosing the distance metric k_1 in Eqn. 4; to be consistent with the final absolute distance metric, we set k_2 to be cosine similarity), KL (88.7%) and L2 (88.8%) exhibit similar performance and outperform L1 (88.3%). Overall, we implement both k and k_2 as ConsineSimilarity and k_1 L2 distance metrics.

4.4.5 Impact of Mix Teacher

Teacher	Ensemble	A	C	P	S	Avg
CLIP ViT B/16	No	88.0	85.2	97.8	86.4	89.4
CLIP RN101	No	87.6	86.1	97.6	85.1	89.1
Mix Teacher	Yes	88.7	86.7	98.3	86.9	90.2
CLIP ViT B/16	Yes	88.3	86.0	98.1	86.7	89.8

Table 6. Results of ERM + Hint + AD + RD with different CLIP teachers on PACS. Hint [19] stands for the distillation loss. AD stands for absolute distance loss. RD stands for relative distance loss. MT stands for Mix Teacher engineering technique. A, C, P, and S: art-painting, cartoon, photo, and sketch.

Finally, we explore the impact of having multiple CLIP teachers, which we call "Mix Teacher" (MT). Specifically, we use another CLIP ResNet 101 as a teacher model, which achieves 94.9%, 80.0%, and 76.0% zero-shot inference on PACS, VLCS, and Office-Home, respectively, and 50.5% finetune inference on TerraIncognita. Our ERM + Hint + AD + RD method with this CLIP RN101 teacher achieves 89.1%, 81.6%, 70.9% and 52.3% on PACS, VLCS, Office-Home and Terra respectively.

Table 6 shows the ensembling results. Overall, an ensemble of teachers achieves higher accuracy compared to non-ensemble teachers; from Table 6, we see that Mix Teacher of CLIP ViT B/16 + CLIP RN101 (90.2%) exhibits better performance than CLIP ViT B/16 non-ensemble (89.4%) and CLIP RN101 (89.1%). We also investigate ensembling the outputs of two separate students trained with the same CLIP ViT B/16 teacher (Table 6 last row). This ensemble model also does better (+0.4%) than a single student (1st row), but not as well as ensembling multiple teachers.

Although the overall performance is close between students distilled by different CLIP teachers (row 1 vs row 2 in Table 6), upon closer inspection, we find that the student distilled with CLIP VIT outperforms the CLIP RN counterpart on sketch domains and worse on cartoon domains. We suspect that the teacher CLIP with different model architectures have different domain biases and perform well on different domains. Thus, by ensembling student models distilled with different CLIP teachers that have different network architectures, we can further improve the generalization capability of our student method. We report the ensemble performance for two different ResNet50 pretrained student models with mixed teachers in the last row in Table 1. It provides a +1.0% boost (74.8% average accuracy) over our single non-ensemble model.

5. Conclusion

One of the challenges in domain generalization is that machine learning models tend to learn domain-specific features, which can make them less effective at generalizing to new domains. This is because domain-specific features may be highly relevant to the training data but may not be useful for prediction on new domains. To address this challenge, the community has focused on developing methods to regularize the learned representations to become domain-invariant features.

In this paper, we build upon this direction by investigating methods for learning domain-invariant features in machine learning models. Our proposed method is inspired by the intuition that while an image tends to convey rich but sometimes excessive details through its pixels, a corresponding text description can describe the crux of the image content in a highly concise and complementary manner.

Following this intuition, we proposed RISE, with main loss functions (the absolute distance and relative distance) to offer specific guidance on how the student model's training process should be regularized, that provides a powerful new direction for domain generalization research by incorporating the power of language to regularize image representations.

Our results suggest that leveraging language as a regularization strategy can achieve state-of-the-art performance on popular domain generalization benchmarks. We have demonstrated the effectiveness of our approach through a comprehensive set of experiments.

In conclusion, RISE provides a new direction for domain generalization research by incorporating the power of language to regularize image representations. Our results suggest that leveraging language as a regularization strategy can significantly improve the generalization capability of machine learning models, and we believe that our work can motivate further research in this direction.

Limitations. When facing the downstream task, such as Terra where CLIP shows poor performance during zero-shot inference, finetuning CLIP on the downstream task is recommended before distilling knowledge to students.

In addition, the quality and relevance of text descriptions used to regularize image representations may impact the effectiveness of our approach. In our experiments, we addressed this limitation by using an average of 80 differ-

ent text descriptions for each image. However, obtaining a more direct and generic text description might help improve the efficiency of the method. We leave this to future work.

Acknowledgements. This work was supported in part by NSF CAREER IIS2150012, NASA 80NSSC21K0295, and Institute of Information & Communications Technology Planning & Evaluation(IITP) grants funded by the Korean government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training.

References

[1] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. *arXiv preprint arXiv:1904.12543*, 2019. 1, 2

[2] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint arXiv:2110.10832*, 2021. 6

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 5

[4] Fabio M Carlucci, Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Agnostic domain generalization. *arXiv preprint arXiv:1808.01102*, 2018. 1, 2

[5] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *NeurIPS*, 2021. 6

[6] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV*, 2022. 2

[7] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *IEEE Trans. Neural Networks Learn. Syst.*, 32(1):25–35, 2021. 3

[8] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018. 3

[9] Yu Ding, Lei Wang, Bin Liang, Shuming Liang, Yang Wang, and Fang Chen. Domain generalization by learning and removing domain-specific features. *arXiv preprint arXiv:2212.07101*, 2022. 2

[10] Songwei Ge, Haohan Wang, Amir Alavi, Eric Xing, and Ziv Bar-Joseph. Supervised adversarial alignment of single-cell rna-seq data. *Journal of Computational Biology*, 2021. 1, 2

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019. 2

[13] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 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] Yushuo Guan, Pengyu Zhao, Bingxuan Wang, Yuanxing Zhang, Cong Yao, Kaigui Bian, and Jian Tang. Differentiable feature aggregation search for knowledge distillation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 3

[16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 5, 6

[17] Beining Han, Chongyi Zheng, Harris Chan, Keiran Paster, Michael R Zhang, and Jimmy Ba. Learning domain invariant representations in goal-conditioned block mdps. *arXiv preprint arXiv:2110.14248*, 2021. 1, 2

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 3

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6, 7, 8

[20] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 3

[21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *ICCV*, 2021. 2

[22] Zeyi Huang, Haohan Wang, Dong Huang, Yong Jae Lee, and Eric P Xing. The two dimensions of worst-case training and their integrated effect for out-of-domain generalization. In *CVPR*, 2022. 2, 6

[23] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 2

[24] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *CoRR*, abs/1707.01219, 2017. 3

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 3

[26] Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS*, 2018. 3

[27] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross-domain ensemble distillation for domain generalization. In *ECCV*, 2022. 2

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 5

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 1, 2, 6

[30] Yante Li, Wei Peng, and Guoying Zhao. Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation. In *FG*, 2021. 3

[31] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 1, 2

[32] Ziyue Li, Kan Ren, Xinyang Jiang, Bo Li, Haipeng Zhang, and Dongsheng Li. Domain generalization using pretrained models without fine-tuning. *arXiv preprint arXiv:2203.04600*, 2022. 2

[33] Junjie Liu, Dongchao Wen, Hongxing Gao, Wei Tao, Tse-Wei Chen, Kinya Osa, and Masami Kato. Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation. In *CVPR Workshops*, 2019. 3

[34] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *CVPR*, 2019. 3

[35] Rang Meng, Xianfeng Li, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Mingli Song, Di Xie, and Shiliang Pu. Attention diversification for domain generalization. In *ECCV*, 2022. 2

[36] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 1, 2

[37] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 1, 2

[38] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, 2021. 6

[39] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atılım Güneş Baydin. Domain invariant representation learning with domain density transformations. *arXiv preprint arXiv:2102.05082*, 2021. 1, 2

[40] Mauricio Orbes-Arteainst, Jorge Cardoso, Lauge Sørensen, Christian Igel, Sebastien Ourselin, Marc Modat, Mads Nielsen, and Akshay Pai. Knowledge distillation for semi-supervised domain adaptation. In *International Workshop on OR and MLCN*, 2019. 3

[41] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 3

[42] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *CVPR*, 2020. 3

[43] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Trans. Neural Networks Learn. Syst.*, 32(5):2030–2039, 2021. 3

[44] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. ALP-KD: attention-based layer projection for knowledge distillation. In *AAAI*, 2021. 3

[45] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. In *ICCV*, 2019. 3

[46] 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 *ICML*, 2021. 2, 3, 4, 7

[47] Mohammad Mahfujur Rahman, Clinton Fookes, and Sridha Sridharan. Discriminative domain-invariant adversarial network for deep domain generalization. *arXiv preprint arXiv:2108.08995*, 2021. 1, 2

[48] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, 2020. 6

[49] 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. 2

[50] Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In *AAAI*, 2019. 3

[51] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 2

[52] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 6

[53] Chris Xing Tian, Haoliang Li, Xiaofei Xie, Yang Liu, and Shiqi Wang. Neuron coverage-guided domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[54] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 5

[55] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 5

[56] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. 2017. 1, 2

[57] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[58] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *CVPR*, 2022. 2

[59] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*. 3

[60] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *ICM*, 2021. 3

[61] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *CVPR*, 2022. 2

[62] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Lili Ju, and Song Wang. Siamdoge: Domain generalizable semantic segmentation using siamese network. In *ECCV*, 2022. 2

[63] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, 2020. 3

[64] Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 3

[65] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, 2022. 3

[66] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Lu Yuan, Ce Liu, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *CVPR*, 2022. 3

[67] Jing Yang, Brais Martínez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via adaptive instance normalization. *CoRR*, abs/2003.04289, 2020. 3

[68] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 3

[69] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022. 2

[70] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*, 2022. 6

[71] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021. 5

[72] Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 3

[73] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, 2022. 3

[74] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *CVPR*, 2019. 3

[75] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019. 2

[76] Chenrui Zhang and Yuxin Peng. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In Jérôme Lang, editor, *IJCAI*, 2018. 3

[77] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Mvdg: A unified multi-view framework for domain generalization. In *ECCV*, 2022. 2

[78] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 3

[79] Xin Zhang, Shixiang Shane Gu, Yutaka Matsuo, and Yusuke Iwasawa. Domain prompt learning for efficiently adapting clip to unseen domains. *arXiv e-prints*, pages arXiv–2111, 2021. 2, 6

[80] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *NeurIPS*, 2020. 1, 2

[81] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3

[82] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020. 2

[83] Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *CVPR*, 2022. 2