PCL: Proxy-based Contrastive Learning for Domain Generalization

Xufeng Yao†, Yang Bai†, Xinyun Zhang†, Yuechen Zhang†, Qi Sun†, Ran Chen†, Ruiyu Li‡, Bei Yu†
†The Chinese University of Hong Kong ‡SmartMore
{xfyao,byu}@cse.cuhk.edu.hk

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.00696

Abstract

Domain generalization refers to the problem of training a model from a collection of different source domains that can directly generalize to the unseen target domains. A promising solution is contrastive learning, which attempts to learn domain-invariant representations by exploiting rich semantic relations among sample-to-sample pairs from different domains. A simple approach is to pull positive sample pairs from different domains closer while pushing other negative pairs further apart. In this paper, we find that directly applying contrastive-based methods (e.g., supervised contrastive learning) are not effective in domain generalization. We argue that aligning positive sample-to-sample pairs tends to hinder the model generalization due to the significant distribution gaps between different domains. To address this issue, we propose a novel proxy-based contrastive learning method, which replaces the original sample-to-sample relations with proxy-to-sample relations, significantly alleviating the positive alignment issue. Experiments on the four standard benchmarks demonstrate the effectiveness of the proposed method. Furthermore, we also consider a more complex scenario where no ImageNet pre-trained models are provided. Our method consistently shows better performance.

1. Introduction

Deep neural networks (DNNs) have achieved significant success in various applications, assuming the training and test data are independent and identically distributed (i.i.d.) [2, 12, 19, 20, 25, 40, 41, 47, 52]. However, in many real-world problems, training and testing datasets are collected under different scenarios, which leads to the DNNs trained on the source data performing poorly on the out-of-distribution target data. Such performance degeneration due to domain shift impairs the generalization ability of DNNs. The literature in domain generalization (DG) aims to address this issue by exploiting the diversity of source domains to improve model generalization.

Different from domain adaptation task, it is assumed that only source domains can be accessed during training. Therefore, most prior works in DG task focus on learning domain-invariant representations by aligning different source domains [31, 33, 35]. Contrastive learning provides a potential solution to address this problem. The key idea is to construct multiple positive and negative pairs, which are then used to learn to optimize a distance metric that

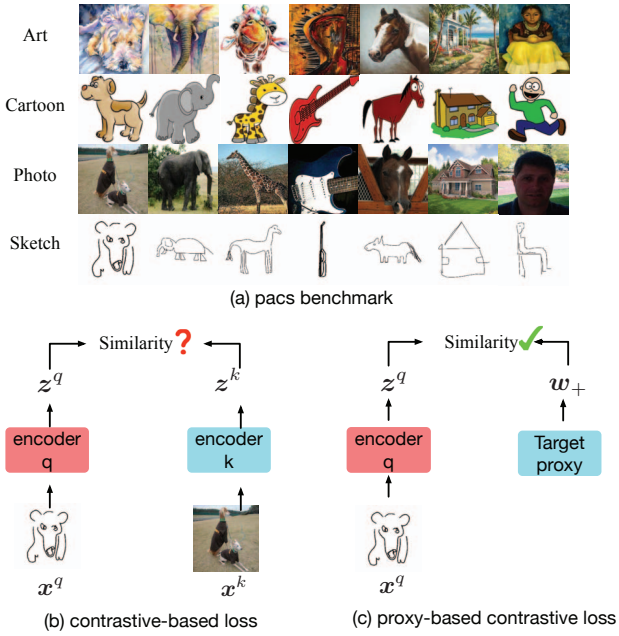
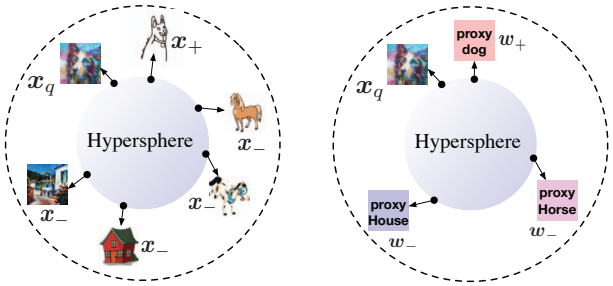


Figure 1. (a) PACS dataset is a typical domain generalization benchmark that contains four domains: Art, Cartoon, Photo and Sketch with seven categories in each domain. Domain Generalization task aims to train the model from multi-source domains (e.g., art, photo, sketch) and test on target domain (e.g., cartoon). In the training stage, the target dataset can not be accessed. (b) Typical contrastive-based loss (e.g., supervised contrastive loss) exploits sample-to-sample relations, where different domain samples from the same class can be regarded as positive pairs. We argue that optimizing some hard positive pairs may worsen the model generalization. We term it as the positive alignment problem. (c) Based on our observation, we propose a proxy-based contrastive loss. By replacing the sample-to-sample relations with proxy-to-sample relations, we largely alleviate the positive alignment problem.

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.00696



(a) Contrastive-based Loss **(b)** Proxy-based Loss

Figure 2. Comparison between contrastive-based loss and proxy-based loss. $\boldsymbol{x}$ means the sample embedding, and $\boldsymbol{w}$ indicates the class proxy. $+$ and $-$ indicate the positive samples and negative samples with respect to the anchor sample $\boldsymbol{x}_q$. All embeddings are normalized to a unit hypersphere. (a) Contrastive-based loss mainly focuses on exploiting dense sample-to-sample relations, e.g., $(\boldsymbol{x}_q, \boldsymbol{x})$. (b) Proxy-based loss uses proxies to represent classes. Typically, proxy-to-sample relations e.g, $(\boldsymbol{x}_q, \boldsymbol{w})$ are far more easier to optimize than sample-to-sample relations with a low training complexity.

brings the positive pairs closer while pushing the negative pairs away. By optimizing the contrastive-based objective, the network can learn generalized features by utilizing rich sample-to-sample relations from different domains.

In this paper, we find that some conventional contrastive-based methods (e.g., supervised contrastive learning) are not effective for domain generalization. One potential reason is that complex positive sample-to-sample relations hamper the model generalization. A good view is very important in contrastive-based loss [53] as too easy or too difficult sample pairs both hinder the model performance. As shown in Figure 1, in a supervised contrastive learning setting, we sample positive sample-to-sample pairs from different domains. However, some positive pairs are very difficult to align due to large domain gaps, which degrades the model generalization.

We attempt to address the problem from a proxy-based approach. A proxy can be regarded as the representative of a sub-dataset, then ideally more robust to the noise samples or outliers. A standard proxy-based approach is softmax CE loss (i.e., softmax cross entropy loss), where the proxies are used to represent classes. The major difference between contrastive-based methods and proxy-based methods is relation construction. As illustrated in Figure 2, contrastive-based loss mainly focuses on exploring rich sample-to-sample relations, while proxy-based loss use proxies to represent sub-trainset, enables safe and fast convergence but missing some semantic relations. This motivates us to design a proxy-based loss that takes some good points from contrastive learning. We regard each proxy as the anchor and consider all proxy-to-sample relations. To prevent model from getting stuck in some trivial solutions, we align a projection head on both sample embeddings and proxy weights and use the new embeddings and new proxy weights for proxy-based contrastive loss.

Our contributions are as follows: First, we empirically unveil the degradation of model generalization from positive alignment problem in contrastive learning. Second, we propose a novel proxy-based contrastive learning technique for domain generalization. The proposed technique is fairly simple yet effective. Third, the proposed algorithm achieves state-of-the-art accuracy on multiple standard benchmarks and consistently improves the model performance in a more complex scenario where ImageNet pre-trained models are not provided.

2. Related Work

Domain Generalization. Most domain generalization methods can be categorized into three groups: (1) Data Augmentation is commonly used to improve the generalization of DNNs as a regularization approach. [37, 56, 70] mainly focus on image-level data manipulation, while [72] proposes a feature-level augmentation on DG. (2) Learning Strategy: Some works attempt to promote the model generalization through learning strategies such as ensemble-learning and meta-learning. Ensemble learning assumes each domain contains some domain-specific knowledge that can be learned better by different networks. Then these networks can be combined to improve model generalization. Meta-learning aims to learn from episodes sampled from related tasks to benefit future learning. [3, 28, 30, 71] all provide a promising solution for DG. (3) Domain-invariant representation learning: Most works in DG focus on learning domain-invariant representations. One common approach is to minimize some statistical metrics, such as MMD [31] and Wasserstein distance [69]. Another popular method uses adversarial learning [33, 46] to align different source domains. Additionally, some works also propose contrastive-based methods in DG. For instance, EISNet [58] utilizes both self-contrastive and supervised contrastive learning with negative mining in DG. PDEN [32] applies contrastive learning for single domain generalization. SelfReg [23] dives into positive pairs alignment in a self-contrastive manner.

Contrastive Learning. Our work is highly motivated by recent progress in contrastive learning so we also introduce it here. To prevent model from getting stuck in a trivial solution, Moca [18] proposes a momentum update strategy, SimCLR [10] instead adds a projection head to prevent model collapse. Later people find that a simple stop-gradient operation [11, 15] can address the problem easily even without memory-bank or large batch-size [10, 63] to provide large amounts of negative pairs. While some works [15, 66] achieve impressive results without negative pairs. In practice, large amounts of negative pairs guarantee model performance in relatively simple architectures.

Table 1. Positive alignment loss on PACS benchmark

loss function	acc
softmax CE loss	**88.1**
softmax CE loss w. positive align	86.7

On the other hand, contrastive learning that leverages label information also shows great success in many research fields [16, 22].

Some analysis on the behavior of contrastive learning is also inspiring, [60] proposes two important properties in contrastive learning, named alignment and uniformity. [53] analyses the influence of different data augmentation in contrastive and presents the model can not benefit from either too weak or too strong augmentation strategies.

Metric Learning. Our work is also inspired by metric learning. Pair-based losses and proxy-based losses are two main branches of metric learning. Pair-based methods, such as [17, 43, 48, 61], focus on sample-to-sample relations. From this view, contrastive learning can be regarded as a branch of metric learning. On the other hand, proxy-based methods such as Proxy-NCA [34] and NormFace [57], focus on proxy-to-sample relations. Circle Loss [50] proposes a unified metric learning loss function. Proxy-based methods can be seen as a generalization approach, which achieves better generalization with low training complexity with the sacrifice of exploring potential semantic information contained in sample-to-sample relations.

3. Method

3.1. Motivation

We start by motivating our method before introducing its details. We empirically find that some conventional contrastive-based approaches do not contribute to domain generalization task, so we conjecture that there exists a positive alignment problem where complex pairs may hamper the model generalization. Since most contrastive-based losses consider both positive pairs and negative pairs, we first introduce a loss function inspired by [61] that only considers multi-positive sample-to-sample pairs to verify our hypothesis. Assume $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are sampled from different source domains in the same class. Let $\boldsymbol{z} = F_\theta(\boldsymbol{x})$ be the features extracted by the feature extractor F_θ, we have:

$$\mathcal{L}_{\text{pos}} = \frac{1}{\alpha}\log(1 + \sum \exp(-\boldsymbol{z}_i^\top \boldsymbol{z}_j \cdot \alpha)), \tag{1}$$

where $\boldsymbol{z}_i, \boldsymbol{z}_j$ are two normalized embeddings, α is the scale factor. We use a simple softmax CE loss as our baseline.

We use a classical DG benchmark PACS to validate our hypothesis. From Table 1, we can observe that the proposed positive alignment objective does not contribute to the performance, which motivates us to design a novel loss function to address the issue. More details can be found in Section 4.

3.2. Problem Formulation

Domain generalization aims to train a model that can generalize to the unseen target domains by utilizing multiple source domains. The source domains and target domains $\mathcal{D} = \{D_1, D_2, ...D_K\}$ share a common label space. In each domain, samples are drawn from a dataset $D_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_t}$ where N_t is the number of labeled samples in the domain D_k. Our goal is to learn a generalized model G from a collection of source datasets that performs well on target data. We consider an object recognition model composed of a feature extractor, $F_\theta : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a feature embedding space and a classifier $G_\psi : \mathcal{Z} \to \mathbb{R}^C$, where C denotes the number of classes in the label space.

To introduce our method, we first review softmax CE loss and contrastive-based loss. Then we analyze their properties from different angles. Based on the analysis, we present our proxy-based contrastive loss.

3.3. Review Softmax CE loss

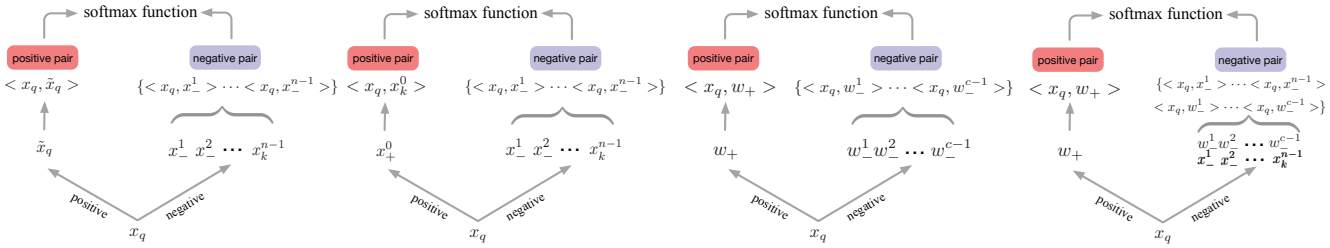
First we review softmax CE loss. Formally, we define the vectors of proxies as $\boldsymbol{w}_i$ $(i = 1, 2, \ldots, C)$ in final FC layer to represent each target class. Here C is the number of the classes. We define $\boldsymbol{z}$ as the feature embedding generated by the feature extractor F_θ, the softmax CE loss can be given by:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{z}_i)}{\exp(\boldsymbol{w}_c^\top \boldsymbol{z}_i) + \sum_{j=1}^{C-1} \exp(\boldsymbol{w}_j^\top \boldsymbol{z}_i)}, \tag{2}$$

where $\boldsymbol{w}_c$ represents the target class. In softmax CE loss, we associate each anchor sample $\boldsymbol{x}$ with all class proxies and measure their similarities with softmax function. Softmax CE loss is an efficient way to learn class proxies with low training complexity i.e., $\mathcal{O}(CN)$. Given C classes, traditional SVM [51] needs to learn $\frac{C(C-1)}{2}$ classifiers using one-vs-one strategy while C SVMs will be trained independently using one-vs-all strategy. In contrast, softmax CE loss only needs to learn C classifiers by aligning a proxy weight (i.e., $\boldsymbol{w}$) for each class, the decision boundary for each pair of class is $(\boldsymbol{w}_i - \boldsymbol{w}_j)^\top \boldsymbol{x} = 0$, where i, j are class indexes.

3.4. Review contrastive-based loss

Though softmax CE loss is efficient, one bottleneck is that it only considers proxy-to-sample relations. Therefore, it ignores rich semantic sample-to-sample relations. In contrast, Contrastive-based loss considers rich sample-to-sample relations. The key idea is to learn a distance that pulls the positive pairs closer and pushes negative pairs apart. To simplify the problem, we do not consider the



(a) self-contrastive loss (b) supervised-contrastive loss (c) softmax loss (d) proxy-based contrastive loss

Figure 3. (a) Self-contrastive loss samples the positive pairs from different views of the same sample and computes the similarity matrix in a contrastive manner. (b) Supervised-contrastive loss constructs the positive pairs from different samples belonging to the same class. (c) Softmax loss only considers proxy-to-sample relations and constructs positive pairs and negative pairs based on each sample. (d) Proxy-based contrastive loss associates each proxy with all data samples, and thus introduces large amounts of negative samples.

multi-positive scenario here. Assuming a mini-batch containing $N+1$ samples, each anchor sample $\boldsymbol{x}$ can be associated to N other samples where the sample pairs from same class are recognized as positive pairs, a contrastive-based loss can be given by:

$$\mathcal{L}_{\mathrm{CL}} = -\log \frac{\exp(\boldsymbol{z}_i^\top \boldsymbol{z}_+ \cdot \alpha)}{\exp(\boldsymbol{z}_i^\top \boldsymbol{z}_+ \cdot \alpha) + \sum \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_- \cdot \alpha)}, \tag{3}$$

where α is the scale factor.

Hard pair mining. We define s_p as the positive pair score i.e., $\boldsymbol{z}_i^\top \boldsymbol{z}_+$, s_n as the negative pair score, i.e., $\boldsymbol{z}_i^\top \boldsymbol{z}_-$. Taking scale factor λ into consideration, we can derive Equation (4), where $\max[s_n^j - s_p]_+$ indicates the pair of (s_p, s_n) with the largest distance. Therefore, contrastive-based loss can implicitly conduct hard pair mining by controlling the scale factor. Hard pair plays an import role in contrastive-based loss, which can help the network learn a better decision boundary, as well as preventing the network from getting stuck in trivial solution. By considering a large amount of negative pairs, contrastive-based loss can learn more informative sample embeddings.

$$\begin{aligned}
\mathcal{L}_{\mathrm{CL}} &= \lim_{\alpha\to\infty} \frac{1}{\alpha} - \log\Big(\frac{\exp(\alpha \cdot s_p)}{\exp(\alpha \cdot s_p) + \sum_{j=1}^{N-1} \exp(\alpha \cdot s_n^j)}\Big) \\
&= \lim_{\alpha\to\infty} \frac{1}{\alpha} \log\Big(1 + \sum_{j=1}^{N-1} \exp(\alpha(s_n^j - s_p))\Big) \\
&= \max[s_n^j - s_p]_+.
\end{aligned} \tag{4}$$

Gradients Analysis. As illustrated in Figure 3, we can observe that both softmax CE loss and contrastive-based loss construct positive pairs and negative pairs with different relations and can use softmax function to generate output probabilities. Further, the gradients of the positive score and negative score in loss with softmax function can be given by:

$$\frac{\partial \mathcal{L}}{\partial s_i} = \begin{cases} p_i - 1, & \text{if i is positive;} \\ p_i, & \text{if i is negative.} \end{cases} \tag{5}$$

Given Equation (5), we can further derive that $\sum_{j\neq i} \frac{\partial \mathcal{L}}{\partial s_j} = |\frac{\partial \mathcal{L}}{\partial s_i}|$ as $\sum p_i = 1$ where i represents the positive index and j represents the negative index. The equation demonstrates that by pulling the positive pair closer, the negative pairs are pushed by the same strength. Then the number of negative pairs becomes a double-edged sword. In contrastive-based loss, each positive sample pair has to push other negative sample pairs away, which motivates the positive pair to get a higher score. In domain generalization, some positive pairs are formed from different distributions that are difficult to align, which may impair model performance.

3.5. Proxy-based Contrastive Learning

Softmax loss is efficient in learning class-proxy and enables a fast and safe convergence but does not consider the sample-to-sample relations. Contrastive-based loss utilizes rich sample-to-sample relations, but suffers from the high training complexity for optimizing dense sample-to-sample relations. Thus some complex relations may hinder the performance. It's not trivial to design a novel loss function that takes advantage of both softmax CE loss and contrastive-based loss.

For each anchor sample $\boldsymbol{x}_i$, we associate it with all samples in mini-batch, we ignore the positive pair and only consider the negative pairs. On the other hand, we use target class proxy to form the positive pair with the anchor sample. The proxy-based contrastive loss can be given by:

$$\mathcal{L}_{\mathrm{PCL}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{z}_i \cdot \alpha)}{Z}, \tag{6}$$

where Z is given by:

$$Z = \exp(\boldsymbol{w}_c^\top \boldsymbol{z}_i \cdot \alpha) + \sum_{k=1}^{C-1} \exp(\boldsymbol{w}_k^\top \boldsymbol{z}_j \cdot \alpha) + \sum_{j=1, j\neq i}^{K} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_j \cdot \alpha).$$

Here N is the number of samples in one mini-batch, $\boldsymbol{w}_c$ denotes the target class proxy weight of $\boldsymbol{x}_i$. C is the number of classes, K is the number of negative pairs in all $\boldsymbol{x}_i$-based sample-to-sample relations. Both sample embedding i.e., $\boldsymbol{z}$ and proxy weights i.e., $\boldsymbol{w}$ are normalized. α is the scale factor.

Projection Head. We further consider applying projection head for both sample embedding, i.e., $\boldsymbol{z}$ and proxy weight, i.e., $\boldsymbol{w}$ inspired by [10]. A projection head is a small network that maps the embedding to the space that proxy-based contrastive loss is applied. We use a three-layer MLP $h(\cdot)$ as sample embedding projection head and one-layer MLP $g(\cdot)$ projection head for proxy weight. Thus the new embedding and proxy weight can be given by $\boldsymbol{e}_i = h(\boldsymbol{z}_i)$ and $\boldsymbol{v}_i = g(\boldsymbol{w}_i)$. The motivation of applying the projection head is not trivial. Since proxy-based methods are effortless to get converged, the output of the score function tends to be a sparse matrix, which does not have enough strength to push proxy, i.e., $\boldsymbol{w}$ and sample embedding $\boldsymbol{z}$ to explore more semantic feature. A projection head can map both proxy weight and sample embedding to another space. Then the proxy-based contrastive loss is applied, which is harder to converge than softmax loss. Then both proxy weights and sample embedding can learn more meaningful features by the back-propagation.

In-domain negative pair generation and domain sampling strategy. We also consider an in-domain negative pair generation. As discussed in the previous subsection, hard pairs play an important role in contrastive learning. In practice, some negative pairs which are formed by different domains only contain small values that contribute little to optimizing. Therefore, we only consider in-domain negative pairs. Then we have:

$$\mathcal{L}_{\text{PCL-in}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{v}_c^\top \boldsymbol{e}_i)}{E}, \tag{7}$$

where E is given by:

$$E = \exp(\boldsymbol{v}_c^\top \boldsymbol{e}_i) + \sum_{k=1}^{C-1}\exp(\boldsymbol{v}_k^\top \boldsymbol{e}_j) + \sum_{j=1,j\neq i}^{B}\exp(\boldsymbol{e}_i^\top \boldsymbol{e}_j). \tag{8}$$

Both sample embedding $\boldsymbol{z}$ and proxy weight $\boldsymbol{w}$ go through a project head and produce new sample embedding $\boldsymbol{e}$ and new proxy weight $\boldsymbol{v}$. All parts are equivalent to the previous equation except the negative sample-to-sample pairs which consider sample pairs in the same domain. For the balance of negative pair generation, we also take a balanced domain sampling strategy. In each training iteration, we sample the same number of samples from each source domain, which means for each mini-batch training iteration:

$$N = \sum_{d=1}^{D} B_d.$$

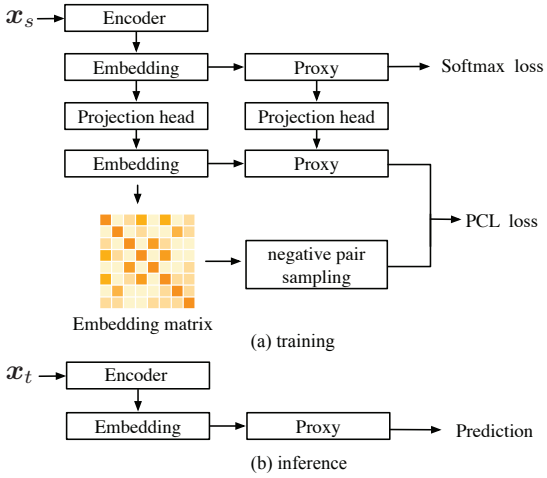


Figure 4. Structure of the proposed proxy-based contrastive loss: (a) Training process; (b) Inference process.

Whole structure. The whole structure is illustrated in Figure 4. In the training stage, we align different projection head for both sample embedding and proxy weight. Then we only select negative pairs from the embedding matrix to construct the proxy-based contrastive loss coupled with proxy weight. The final loss is given by:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{PCL-in}}, \tag{9}$$

where $\mathcal{L}_{CE}$ is simply a softmax CE loss. In inference stage, we only use the original sample embedding and proxy for prediction without introducing extra parameters.

4. Experimental Results

In this section, We first demonstrate the details of the positive alignment experiment that we introduce in the Section 3. Then we evaluate the proposed Proxy-based contrastive loss on four standard DG datasets, PACS [27], Office-Home [55], DomainNet [39] and TerraIncognita [5]. In practice, we use the loss introduced in Equation (7) which means we only use the negative pairs that generate in the same domain. Our work is built on SWAD [8]. For a fair comparison, we follow the same training and evaluation protocol including data splits, hyperparameter search and model selection. We report out-of-domain accuracy for each domain. We also evaluate our method on standard DG benchmarks in a non-pretrained model setting. More details about the results analysis, ablation study, implementation details will be discussed in the ensuing sections.

4.1. Details of the Positive Alignment experiments

We first show the details of the positive alignment experiments that we introduce in the method section. We evaluate the impact of positive alignment on different benchmarks to validate whether it can improve the model generalization.

Table 2. Positive alignment on different benchmarks with ResNet50 ImageNet pre-trained model

Method	PACS	OfficeHome	TerraIncognita
softmax CE	**88.1**	**70.6**	**50.0**
softmax CE w. positive align	86.7	70.1	48.5

Table 3. Comparison with state-of-the-art methods on PACS benchmark with ResNet-50 ImageNet pre-trained model

Algorithm	A	C	P	S	Avg.
IRM [1]	84.8	76.4	96.7	76.1	83.5
MetaReg [4]	87.2	79.2	97.6	70.3	83.6
DANN [14]	86.4	77.4	97.3	73.5	83.7
ERM [54]	85.7	77.1	97.4	76.6	84.2
GroupDRO [42]	83.5	79.1	96.7	78.3	84.4
MTL [6]	87.5	77.1	96.4	77.3	84.6
I-Mixup [62, 64, 65]	86.1	78.9	97.6	75.8	84.6
MMD [31]	86.1	79.4	96.6	76.5	84.7
VREx [26]	86.0	79.1	96.9	77.7	84.9
MLDG [29]	85.5	80.1	97.4	76.6	84.9
ARM [67]	86.8	76.8	97.4	79.3	85.1
RSC [21]	85.4	79.7	97.6	78.2	85.2
Mixstyle [72]	86.8	79.0	96.6	78.5	85.2
ER [68]	87.5	79.3	**98.3**	76.3	85.3
pAdaIN [38]	85.8	81.1	97.2	77.4	85.4
ERM [54]	84.7	80.8	97.2	79.3	85.5
EISNet [59]	86.6	81.5	97.1	78.1	85.8
CORAL [49]	88.3	80.0	97.5	78.8	86.2
SagNet [36]	87.4	80.7	97.1	80.0	86.3
DSON [44]	87.0	80.6	96.0	**82.9**	86.6
SWAD [8]	89.3	83.4	97.3	82.5	88.1
Ours	**90.2**	**83.9**	98.1	82.6	**88.7**

Table 4. Comparison with state-of-the-art methods on OfficeHome benchmark with ResNet-50 ImageNet pre-trained model

Algorithm	**A**	**C**	**P**	**R**	**Avg**
Mixstyle [72]	51.1	53.2	68.2	69.2	60.4
IRM [1]	58.9	52.2	72.1	74.0	64.3
ARM [67]	58.9	51.0	74.1	75.2	64.8
RSC [21]	60.7	51.4	74.8	75.1	65.5
CDANN [31]	61.5	50.4	74.4	76.6	65.7
DANN [14]	59.9	53.0	73.6	76.9	65.9
GroupDRO [42]	60.4	52.7	75.0	76.0	66.0
MMD [31]	60.4	53.3	74.3	77.4	66.4
MTL [6]	61.5	52.4	74.9	76.8	66.4
VREx [26]	60.7	53.0	75.3	76.6	66.4
ERM [54]	61.3	52.4	75.8	76.6	66.5
MLDG [29]	61.5	53.2	75.0	77.5	66.8
ERM [54]	63.1	51.9	77.2	78.1	67.6
I-Mixup [62, 64, 65]	62.4	54.8	76.9	78.3	68.1
SagNet [36]	63.4	54.8	75.8	78.3	68.1
CORAL [49]	65.3	54.4	76.5	78.4	68.7
SWAD [8]	66.1	57.7	78.4	80.2	70.6
Ours	**67.3**	**59.9**	**78.7**	**80.7**	**71.6**

From Table 2, we can observe that the positive alignment objective is not effective for the model generalization on different benchmarks.

Table 5. Comparison with state-of-the-art methods on TerraIncognita benchmark with ResNet-50 ImageNet pre-trained model

Algorithm	**Location100**	**Location38**	**Location43**	**Location46**	**Avg.**
MMD [31]	41.9	34.8	57.0	35.2	42.2
GroupDRO [42]	41.2	38.6	56.7	36.4	43.2
Mixstyle [72]	54.3	34.1	55.9	31.7	44.0
ARM [67]	49.3	38.3	55.8	38.7	45.5
MTL [6]	49.3	39.6	55.6	37.8	45.6
CDANN [31]	47.0	41.3	54.9	39.8	45.8
ERM [54]	49.8	42.1	56.9	35.7	46.1
VREx [26]	48.2	41.7	56.8	38.7	46.4
RSC [21]	50.2	39.2	56.3	40.8	46.6
DANN [14]	51.1	40.6	57.4	37.7	46.7
IRM [1]	54.6	39.8	56.2	39.6	47.6
CORAL [49]	51.6	42.2	57.0	39.8	47.7
MLDG [29]	54.2	44.3	55.6	36.9	47.8
I-Mixup [62, 64, 65]	**59.6**	42.2	55.9	33.9	47.9
SagNet [36]	53.0	43.0	57.9	40.4	48.6
ERM [54]	54.3	42.5	55.6	38.8	47.8
SWAD [8]	55.4	44.9	59.7	39.9	50.0
Ours	58.7	**46.3**	**60.0**	**43.6**	**52.1**

Table 6. Comparison with state-of-the-art methods on DomainNet benchmark with ResNet-50 ImageNet pre-trained model

Algorithm	**clip**	**info**	**paint**	**quick**	**real**	**sketch**	**Avg**
MMD [31]	32.1	11.0	26.8	8.7	32.7	28.9	23.4
GroupDRO [42]	47.2	17.5	33.8	9.3	51.6	40.1	33.3
VREx [26]	47.3	16.0	35.8	10.9	49.6	42.0	33.6
IRM [1]	48.5	15.0	38.3	10.9	48.2	42.3	33.9
Mixstyle [72]	51.9	13.3	37.0	12.3	46.1	43.4	34.0
ARM [67]	49.7	16.3	40.9	9.4	53.4	43.5	35.5
CDANN [31]	54.6	17.3	43.7	12.1	56.2	45.9	38.3
DANN [14]	53.1	18.3	44.2	11.8	55.5	46.8	38.3
RSC [21]	55.0	18.3	44.4	12.2	55.7	47.8	38.9
I-Mixup [62, 64, 65]	55.7	18.5	44.3	12.5	55.8	48.2	39.2
SagNet [36]	57.7	19.0	45.3	12.7	58.1	48.8	40.3
MTL [6]	57.9	18.5	46.0	12.5	59.5	49.2	40.6
ERM [54]	58.1	18.8	46.7	12.2	59.6	49.8	40.9
MLDG [29]	59.1	19.1	45.8	13.4	59.6	50.2	41.2
CORAL [49]	59.2	19.7	46.6	13.4	59.8	50.1	41.5
MetaReg [4]	59.8	**25.6**	50.2	11.5	64.6	50.1	43.6
DMG [9]	65.2	22.2	50.0	15.7	59.6	49.0	43.6
ERM [54]	63.0	21.2	50.1	13.9	63.7	52.0	44.0
SWAD [8]	66.0	22.4	53.5	**16.1**	65.8	55.5	46.5
Ours	**67.9**	24.3	**55.3**	15.7	**66.6**	**56.4**	**47.7**

Table 7. Comparison with the state-of-the-art methods on the Office-Home benchmark with the ResNet-18 backbone

Algorithm	**A**	**C**	**P**	**R**	**Avg**
Deep-All	52.06	46.12	70.45	72.45	60.27
D-SAM [13]	58.03	44.37	69.22	71.45	60.77
Jigen [7]	53.04	47.51	71.47	72.79	61.20
MMD-AAE [31]	56.50	47.30	72.10	74.80	62.70
DSON [45]	59.37	45.70	71.84	74.68	62.90
RSC [21]	58.42	47.90	71.63	74.54	63.12
L2A-OT [70]	60.60	50.10	74.80	77.00	65.60
DAEL [71]	59.40	55.10	74.00	75.70	66.10
SelfReg [23]	**63.60**	53.10	76.90	78.10	67.90
Ours	62.10	**58.22**	**77.38**	77.98	**68.92**

4.2. Datasets

1. PACS contains overall 9991 images and 4 domains: photo, art-painting, cartoon and sketch. **2. DomainNet** is recently proposed by [39], which consists of nearly 0.6 million images of 345 classes distributed among 6 domains - painting, quickdraw, real, clipart, sketch and infograph.

3. Office-Home is a popular benchmark for DG evaluation with four domains of distinct styles: Artistic, Clip-Art, Product and Real-World, and each domain contains images of 65 object categories with around 15,500 images in total.
4. TerraIncognita contains 24788 images, 10 classes and 4 domains.

4.3. Implementation Details

We implement our approach in PyTorch, and fine-tune on the models pre-trained on ImageNet which are all provided by PyTorch model zoo. In non-ImageNet pre-trained setting, we do not use any pre-trained weights. The code is mainly built upon the open-source code of SWAD [8] including its training and evaluation protocol. Following SWAD, we use Adam optimizer with a learning rate of 5e-5. We use the same training strategy and data augmentation methods in SWAD. Note that data augmentation methods in SWAD are very simple. Thus SWAD is a very strong baseline compared with other domain generalization methods.

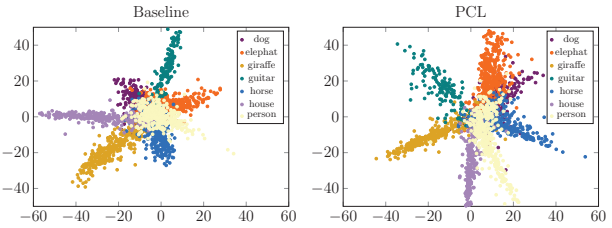
4.4. Results and Discussion

In this section, we evaluate and analyze the results of our approach on four standard benchmarks.

Results on the Domain Generalization benchmarks. For a fair comparison, all models use an identical backbone i.e., ResNet50 and ResNet18 pre-trained on ImageNet. We first compare PCL with the state-of-the-art method, i.e., SWAD. As shown in Tables 3 to 7, our method outperforms SWAD on four benchmarks: OfficeHome, PACS, TerraIncognita and DomainNet. First, our results outperforms some conventional DG methods such as ERM [54], IRM [1] and MMD [31]. Second, our method also surpasses some classical data augmentation methods such as variants of Mixup [64] and Jigen [7]. Third, our performance also beats some ensemble learning approaches such as DAEL [71] and DSON [45]. We also compare our method with the state-of-the-art contrastive learning algorithms such as EISNet [59] and SelfReg [23] which demonstrates the effectiveness of our method. Note that we follow the same data split strategy as in SWAD which splits the data into training (60%), testing (20%) and validation parts (20%). In Table 7, we follow the same data split strategy as SelfReg did.

Comparison with state-of-the-art methods without ImageNet pre-trained. We also want to verify the effectiveness of our method on a non-ImageNet pre-trained setting. First, most domain generalization benchmarks' labels are overlapped with ImageNet's labels, which somehow influence the learning strategy of DG approaches. Second, in some industry AI cases, ImageNet pre-trained models are forbidden to use due to commercial and privacy problems. Thus it's necessary to design a general DG algorithm where no ImageNet pre-trained models are provided. We first validate our method on a small dataset cifar100, which contains 60000 32×32 images in 100 classes, with 600 images per class. We use a simple AlexNet [25] with only one linear layer. The initialized learning rate is set to 0.1 with a step decay schedule. Following is the results:

As shown in Table 8, we use softmax CE loss as baseline. We notice that our method can surpass the softmax CE loss stably.



(a) Softmax (b) PCL

Figure 5. t-SNE visualization on softmax CE loss and proposed PCL Loss. We visualize our embedding on PACS benchmark where the sources are art, photo and sketch, and the target domain is cartoon. We observe our method better captures the domain-invariant features and shows a better results.

Table 8. A Pilot experiment on Cifar100 w/o ImageNet pre-trained

Method	batch-size	top1-acc (%)	top1-error (%)
CE	128	44.11	55.89
CE w. PCL	128	**48.16**	**51.84**
CE	256	46.28	53.72
CE w. PCL	256	**49.12**	**50.88**

Table 9. Comparison with state-of-the-art methods on OfficeHome benchmarks w/o ImageNet pre-trained

Backbone	Method	Art	Clipart	Product	Real-World	avg
ResNet50	SWAD	16.43	18.59	29.84	29.92	23.69
	Ours	**17.04**	**21.94**	**31.81**	**33.39**	**26.05**
ResNet18	SWAD	11.89	15.64	26.55	27.91	20.50
	Ours	**15.50**	**21.08**	**30.46**	**32.04**	**24.77**

Table 10. Comparison with state-of-the-art methods on PACS benchmarks w/o ImageNet pre-trained

Backbone	Method	Art	Cartoon	Photo	Sketch	avg
ResNet50	SWAD	33.80	**54.90**	60.18	42.46	47.84
	Ours	**41.24**	54.42	**60.63**	**49.75**	**51.51**
ResNet18	SWAD	29.04	**50.05**	**52.32**	30.53	40.49
	Ours	**33.56**	47.23	51.65	**44.85**	**44.32**

Table 11. Comparison with state-of-the-art methods on TerraIncognita benchmarks w/o ImageNet pre-trained

	Method	Location100	Location38	Location43	Location46	avg
ResNet50	SWAD	21.01	24.79	**27.80**	23.41	24.25
	Ours	**22.44**	**31.98**	22.42	**23.66**	**25.13**
ResNet18	SWAD	21.49	25.65	19.21	**20.06**	21.60
	Ours	**21.65**	**28.18**	**20.12**	18.16	**22.02**

Table 12. Ablation study on different contrastive-based loss on Office-Home on ResNet18

loss function	avg
softmax CE	66.78
Proxy-Anchor [24]	62.48
softmax CE w. supervised CL	65.98
Ours	**68.92**

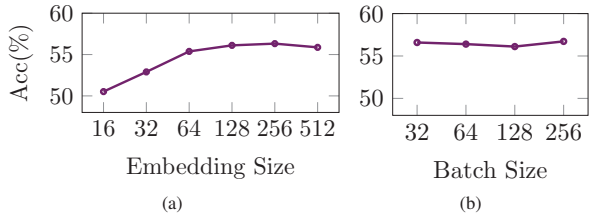


Figure 6. (a) Accuracy versus embedding-size on Office-Home dataset; (b) Ablation study on different batch-size on Office-Home benchmark.

We further consider the model generalization without ImageNet pre-trained on standard DG benchmarks. We test our performance with SWAD on three standard DG benchmarks: OfficeHome, PACS and TerraIncognita. We evaluate the performance on two backbone i.e., ResNet50 and ResNet18. As shown in Tables 9 to 11, our method surpass the SWAD on both ResNet18 and ResNet50 backbone, which demonstrates the effectiveness of our method.

Ablation study on proxy-based and contrastive-based methods. To demonstrate the effectiveness of the proposed Proxy-based pair loss, we compare it with other classical proxy-based losses such as softmax CE Loss and proxy-anchor loss [24]. We also compare our proposed loss function with supervised contrastive loss (supervised CL) [22].

As shown in Table 12, we can find that our method surpasses both proxy-based methods and contrastive-based methods. In particular, we can find that supervised CL loss does not even beat the baseline softmax CE loss in small networks.

Ablation study on embedding size. We conduct an ablation study on Office-Home benchmark where the source domain is art, product and real-world and target domain is clipart. As shown in Section 4.4, we can find that with the increase of embedding size, the model gains more capacity, thus it achieves better results. However, the performance somehow degrades when embedding size is large enough e.g., 128. On the other hand, we can find that our method is still robust to the embedding size, even with a small embedding size like 16. The model can still achieve a comparable result.

Ablation study on batch-size. Batch size is an important metric in contrastive-based loss because it controls the number of sample-to-sample pairs. As demonstrated in Section 4.4, we can observe that the model is stable to batch-size, which is beyond our expectations because large batch size can generate more sample-to-sample pairs. On the other hand, our model is robust to the batch-size even with a small batch size such as 32.

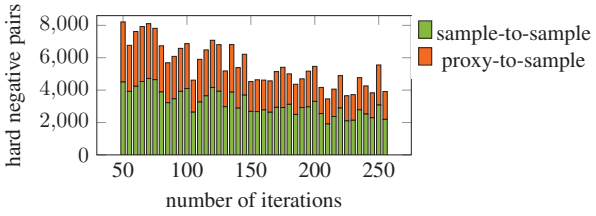


Figure 7. Analysis on hard negative pairs selection.

Effectiveness of negative hard pair selection. Hard negative pair sampling plays a key role in our approach, so we also conduct a statistical analysis on OfficeHome benchmark. We define the hard negative pair as the negative pair whose similarity score is larger than the minimal positive pair's similarity score with a margin $s_n + m \leq \min(s_p)$, and we set m to 0.35 in this work. As illustrated in Figure 7, the sample-to-sample represents the number of hard negative pairs selected from the sample-to-sample pairs e.g., $(\boldsymbol{x}_i, \boldsymbol{x}_-)$. The proxy-to-sample indicates the number of hard negative pairs sampled from the proxy-to-sample pairs e.g., $(\boldsymbol{w}_-, \boldsymbol{x}_j)$. We can find that both sample-to-sample pairs and proxy-to-sample pairs made a stable contribution on hard negative pairs while the number of hard sample-to-sample pairs are much larger than hard proxy-to-sample pair. The total number of negative hard samples becomes smaller during the training process, which means the network has a better feature extraction ability.

5. Conclusion

Limitations. The proxy-based method makes a trade-off between sample-to-sample relations and class-to-sample relations. The model generalization is gained by sacrificing some potential useful semantic relations.

In this paper, we explore the positive alignment problem of contrastive learning in domain generalization. We empirically reveal that the performance degradation in domain generalization of some typical contrastive-based methods stems from the positive pair alignment. Then we introduce a simple yet effective approach, named proxy-based contrastive learning, to address the problem. Our method is easy to implement and performs stably. Without bells and whistles, the proposed method surpasses state-of-the-art methods on several standard DG datasets.

Acknowledgment

This work is partially supported by SmartMore and ITF Partnership Research Programme (No. PRP/65/20FX).

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 6, 7

[2] Yang Bai and Weiqiang Wang. Acpnet:anchor-center based person network for human pose estimation and instance segmentation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1072–1077, 2019. 1

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Conference on Neural Information Processing Systems (NIPS)*, 31:998–1008, 2018. 2

[4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008, 2018. 6

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 5

[6] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22:2–1, 2021. 6

[7] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2229–2238, 2019. 6, 7

[8] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 5, 6, 7

[9] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. 6

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 5

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[13] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, pages 187–198. Springer, 2018. 6

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 6

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[16] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020. 3

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

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

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3, 8

[23] Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. *arXiv preprint arXiv:2104.09841*, 2021. 2, 6, 7

[24] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020. 8

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1, 7

[26] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 6

[27] 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. 5

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2018. 2

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 6

[30] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1446–1455, 2019. 2

[31] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018. 1, 2, 6, 7

[32] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. 2

[33] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 1, 2

[34] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 3

[35] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 1

[36] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 6

[37] Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *European Conference on Computer Vision*, pages 666–670. Springer, 2020. 2

[38] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2021. 6

[39] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019. 5, 6

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[42] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 6

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 3

[44] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020. 6

[45] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision (ECCV)*, pages 68–83. Springer, 2020. 6, 7

[46] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10023–10031, 2019. 2

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 1

[48] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Conference on Neural Information Processing Systems (NIPS)*, pages 1857–1865, 2016. 3

[49] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 6

[50] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6398–6407, 2020. 3

[51] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 3

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 1

[53] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020. 2, 3

[54] V Vapnik. Statistical learning theory new york. *NY: Wiley*, 1:2, 1998. 6, 7

[55] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027, 2017. 5

[56] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018. 2

[57] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 3

[58] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020. 2

[59] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020. 6, 7

[60] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 3

[61] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5022–5030, 2019. 3

[62] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020. 6

[63] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[64] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 6, 7

[65] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 6

[66] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 2

[67] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv preprint arXiv:2007.02931*, 2020. 6

[68] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33, 2020. 6

[69] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573*, 2020. 2

[70] 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. 2, 6

[71] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *arXiv preprint arXiv:2003.07325*, 2020. 2, 6, 7

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