OUT-OF-DISTRIBUTION GENERALIZATION OF SELF-SUPERVISED CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning attracts much attention recently, since it does not require labeled data for training contrasted to supervised learning. Empirical studies observe that it also has better transfer ability than supervised learning. However, the theoretical study of the out-of-distribution (OOD) generalization ability of self-supervised learning is still limited. In this paper, we establish a theoretical framework to analyze the OOD performance of contrastive-based self-supervised learning, by connecting the OOD generalization ability with the data augmentation used in SSL. Although some recent work claims that contrastive learning learns more robust representations than supervised learning, our results suggest that this superiority mainly comes from the data augmentation used, i.e., more data are fed to the model. In the face of more challenging OOD scenarios, the standard contrastive learning still suffers from the same generalization problem as empirical risk minimization (ERM). Based on our theoretical results, we propose an augmentation-robust contrastive learning approach, named as ArCL, which significantly improves the OOD performance of contrastive learning in several datasets.

1 INTRODUCTION

Classical machine learning algorithms commonly assume that training and test samples come from the same distribution, i.e., the IID setting. However, in real-world applications, this assumption can be easily violated, and classical algorithms can perform poorly due to the difference between training and test distribution. This motivates the study of out-of-distribution (OOD) generalization problem, which aims to develop algorithms that can generalize to unseen test distributions (Blanchard et al., 2011; Muandet et al., 2013; Wang et al., 2021a; Shen et al., 2021).

Obviously, generalization to arbitrary distributions is impossible. Different characterization of the connection between test and training distributions leads to distinct algorithms. In the supervised learning regime, most of the proposed approaches aim to find the features that extract some invariance across different distributions, and assume that such invariance also holds in unseen test distributions (Peters et al., 2016; Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Mahajan et al., 2021; Jin et al., 2020; Ye et al., 2021).

For the self-supervised learning (SSL) setting, related works are still limited. Self-supervised learning has attracted great attention in many fields (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021), which firstly learns a representation through a large number of unlabeled training data, and then uses the learned encoder for downstream tasks. Since its two-step nature, self-supervised learning is more likely to meet the OOD issue, and its generalization ability on downstream tasks with different distributions has drawn increasing attention. Some recent works have studied this issue empirically (Liu et al., 2021; Goyal et al., 2021; von Kügelgen et al., 2021; Wang et al., 2021b). However, a thorough theoretical understanding of the OOD generalization ability of contrastive learning is still limited, which hinders the development of tailor-made OOD algorithms.

SSL, especially contrastive-based SSL (contrastive learning), employs the idea that finding representations that are invariant under some data augmentation, which is similar to the invariance-based OOD methods. From this point of view, we wonder whether contrastive learning learns a domain invariant representation.

In this paper, we try to theoretically answer the following questions: (a) on which downstream distributions, the representation obtained by contrastive learning has good generalization performance; (b) whether contrastive learning learns a domain invariant representation. We investigate the first question by relating contrastive loss to downstream risks, and finds the crucial role of data augmentation. For the second question, unfortunately, we discover that contrastive learning fails to learn a domain-invariant representation, even across domains induced by data augmentation. To address this issue, we propose the augmentation-robust contrastive learning (ArCL) approach, a module that can be applied to various popular contrastive learning algorithms, such as SimCLR and MoCo. It forces the learned representation to align positive samples uniformly, and thus can improve its OOD generalization performance with theoretical guarantee. We conduct several experiments on CIFAR10 and ImageNet to show the performance of our method.

Related Work

OOD generalization for supervised learning. OOD generalization, has drawn much attention recently (Blanchard et al., 2011; Muandet et al., 2013; Wang et al., 2021a; Shen et al., 2021). Most of the works focus on learning a representation that performs well on different source domains simultaneously (Rojas-Carulla et al., 2018; Mahajan et al., 2021; Jin et al., 2020), which follows the idea of causal invariance (Peters et al., 2016; Arjovsky et al., 2019). To obtain theoretical guarantees for these methods, structural equation models are often assumed (von Kügelgen et al., 2021; Liu et al., 2020; Mahajan et al., 2021).

There are also many other approaches. Distributionally robust optimization optimize a model's worst-case performance over some uncertainty set directly (Krueger et al., 2021; Sagawa et al., 2019; Duchi & Namkoong, 2021; Duchi et al., 2021). Stable learning (Shen et al., 2020; Kuang et al., 2020) learns a set of global sample weights that could remove the confounding bias for all the potential treatments from data distribution. Disentangled representation learning (Bengio et al., 2013; Träuble et al., 2021; Kim & Mnih, 2018) aims to learn representations where distinct and informative factors of variations in data are separated.

Contrastive self-supervised learning. Early works such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020) use the InfoNCE loss to pull the positive samples close while enforcing them away from the negative samples in the embedding space. These methods require large batch sizes or carefully designed negative sampling strategies. To obviate these issues, some recent works get rid of negative samples and prevent representation collapse by cross-correlation loss (Zbontar et al., 2021) or optimization tricks (Grill et al., 2020; Chen & He, 2021).

A number of recent works also aim to theoretically explain the success of contrastive learning in IID settings. One way to explain it is through the mutual information between positive samples (Tian et al., 2020; Hjelm et al., 2018; Tschannen et al., 2019). Arora et al. (2019) directly analyze the generalization of InfoNCE loss based on the assumption that positive samples are drawn from the same latent classes. Wang & Isola (2020) studies the behavior of InfoNCE loss from the perspective of alignment and uniformity. Huang et al. (2022) take data augmentation into account, and provide generalization bounds for k-nearest neighbor error.

The study of contrastive learning in domain shift setting is still limited. (Shen et al., 2022) and (HaoChen et al., 2022) study contrastive learning in unsupervised domain adaptation setting, where unlabeled target data are obtained. Other robust contrastive learning methods (Kim et al., 2020; Jiang et al., 2020) propose mainly focus on the *adversarial robustness*, not the *distributional robustness* studied in this paper.

2 Problem Formulation

Given a set of unlabeled data where each sample X is i.i.d. sampled from training data distribution $\mathcal{D}$ on $\mathcal{X} \subseteq \mathbb{R}^d$, the goal of Self-Supervised Learning (SSL) is to learn a general encoder $f\colon \mathcal{X} \to \mathbb{R}^m$ for different downstream tasks. Contrastive learning is a popular approach of SSL, which augments each sample X twice to obtain a positive pair (X_1, X_2), and then learns the encoder f by pulling them close and pushing random samples (also called negative samples) away in the embedding space. The data augmentation is done by applying a transformation A to the original data, where A is randomly selected from a transformation set $\mathcal{A}$ according to the distribution π. We use $\mathcal{D}_A$ to

denote the distribution of augmented data under a specific augmentation A, and use $\mathcal{D}_\pi$ to denote the distribution of augmented data, i.e., $\mathcal{D}_\pi = \int_A \mathcal{D}_A \, \mathrm{d}\pi(A)$. The loss of contrastive learning typically consists of two terms (Huang et al., 2022):

$$\mathcal{L}_{\mathrm{con}}(f; \mathcal{D}, \pi) := \mathcal{L}_{\mathrm{align}}(f; \mathcal{D}, \pi) + \lambda \mathcal{L}_{\mathrm{reg}}(f; \mathcal{D}, \pi), \tag{1}$$

where $\mathcal{L}_{\mathrm{align}}(f; \mathcal{D}, \pi) := \mathbb{E}_{X \sim \mathcal{D}} \mathbb{E}_{(A_1, A_2) \sim \pi^2} \|f(A_1(X)) - f(A_2(X))\|^2$ measures the alignment of positive samples, and $\mathcal{L}_{\mathrm{reg}}(f; \mathcal{D}, \pi)$ is the regularization term to avoid feature collapse. For example, the regularization term in InfoNCE loss is

$$\mathcal{L}_{\mathrm{reg}}(f; \mathcal{D}, \pi) := \mathop{\mathbb{E}}_{(X, X') \sim \mathcal{D}^2} \mathop{\mathbb{E}}_{(A_1, A_2, A') \sim \pi^3} \log \left(e^{f(A_1(X))^\top f(A_2(X))} + e^{f(A_1(X))^\top f(A'(X'))} \right).$$

When subscript is omitted, notation $\|\cdot\|$ stands for L^2-norm or Frobenius norm for vectors and matrices in this paper. The encoder f is obtained by minimizing $L_{\mathrm{con}}(f; \mathcal{D}, \pi)$. For simplicity, we assume that $\|f(x)\| = 1$ for every $x \in \mathcal{X}$ and f is L-Lipschitz continuous throughout this paper.

For a downstream task with target distribution $\mathcal{D}^{\mathrm{tar}}$, where the labels are obtained, the learned encoder f will be used. In this paper, we consider a multi-class classification problem $\mathcal{X} \subseteq \mathbb{R}^d \to \mathcal{Y} = \{1, \ldots, K\}$. We focus on the covariate shift setting, i.e., we allow $\mathcal{D}^{\mathrm{tar}}$ to be different from $\mathcal{D}$ but $P(Y|X)$ is fixed. We use the best linear classifier on top of f for classification, and use its risk to evaluate the performance of f over $\mathcal{D}^{\mathrm{tar}}$. Namely, we define

$$\mathcal{R}(f; \mathcal{D}^{\mathrm{tar}}) := \min_{h \in \mathbb{R}^{K \times m}} \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{tar}}} \ell(h \circ f(X), Y) \tag{2}$$

as the downstream risk of representation f on distribution $\mathcal{D}^{\mathrm{tar}}$, where ℓ is some loss function. We focus on the case where ℓ is the square loss. With a slight abuse of notation, for the classifier $h \circ f : \mathbb{R}^m \to \mathbb{R}^k$ we also use $\mathcal{R}(h \circ f; \mathcal{D}^{\mathrm{tar}})$ to denote its risk. In this paper, we study the generalization ability of representation learned by contrastive learning on downstream tasks with different distributions.

At the end of this section, we remark the difference between transformations and augmentations. Transformations are deterministic mapping from $\mathcal{X}$ to $\mathcal{X}$, and augmentations are random application of them. The purpose of this hierarchical model is as follows: (a) even $\mathcal{A}$ is fixed, the choice of augmentation π is crucial to both the IID and OOD performance of contrastive learning; (b) most augmentation methods used in practical indeed satisfy this model. The former we will see in the following theoretical analysis, and here we give examples to demonstrate the latter.

Example 2.1. *The augmentations used in SimCLR (Chen et al., 2020) are compositions of random transformations, such as RandomCrop, HorizonalFlip and Color distortion. Each A denotes a specific composition here. Users will determine probabilities that each transformation will be applied. If used, the parameters of the random transformation such as crop range are also selected randomly. Until the parameters of each used transformations are set, the composited transformation is deterministic. Therefore, we can view these augmentation processes as randomly selecting some deterministic transformations, following the distribution π determined by users.*

Example 2.2. *Some recent works (Jahanian et al., 2021) propose to augment data by transformations in the latent space. Suppose we have a pre-trained generative model with an encoder $T : \mathcal{X} \to \mathcal{Z}$ and a generator $G : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{Z}$ is a latent space. Since $\mathcal{Z}$ is usually simple, one can augment data in $\mathcal{X}$ by randomly shifting $T(X)$. In this setting, the augmentation can be parameterized by $A_\theta(X) = G(T(X) + \theta)$ for $\theta \in \mathcal{Z}$. The distribution π of θ is usually chosen to be a normal distribution.*

3 Out-of-distribution Analysis: the Crucial Role of Data Augmentation

In this section, we theoretically demonstrate that data augmentation plays a crucial role in the superior OOD performance of contrastive learning.

The first step is to establish the connection between contrastive loss $\mathcal{L}_{\mathrm{CL}}(f)$ and downstream risk $\mathcal{R}(f)$. For this purpose, we introduce the (σ, δ)-augmentation notion proposed in (Huang et al., 2022), which was used there to study the KNN error of contrastive learning.

Definition 3.1 ((σ,δ)-transformation)**.** *Let $C_k \subseteq \mathcal{X}$ be the set of all points in class k. A data transformation set $\mathcal{A}$ is called a (σ,δ)-transformation on $\mathcal{D}$ for some $\sigma \in (0,1]$ and $\delta > 0$, if for every $k \in [K]$, there exists $C_k^0 \subseteq C_k$ such that*

$$\mathbb{P}_{X\sim\mathcal{D}}(X \in C_k^0) \geq \sigma\, \mathbb{P}_{X\sim\mathcal{D}}(X \in C_k), \text{ and } \sup_{x_1,x_2\in C_k^0} d_A(x_1,x_2) \leq \delta,$$

where $d_{\mathcal{A}}(x_1,x_2) := \inf_{A_1,A_2\in\mathcal{A}} d(A_1(x_1),A_2(x_2))$ for some distance $d(\cdot,\cdot)$.

This definition measures the diversity of data augmentation. A transformation set with smaller δ and larger σ clusters original data more, i.e., samples from the same class are closer after augmentation. Therefore, one can expect the learned representation f to have better cluster performance, which was measured by KNN error in (Huang et al., 2022). We further extend their KNN result to the downstream risk by taking into account the linear classifier on top of f, and obtain the following lemma.

Lemma 3.1. *Define $\mu_k(f;\mathcal{D}) := \mathbb{E}_{X\sim\mathcal{D}} f(X)$ for $k \in [K]$. Suppose that $\mathcal{A}$ is a (σ,δ)-augmentation. For any representation $f : \mathcal{X} \to \mathbb{R}^{d_1}$ and linear layer $h \in R^{K\times d_1}$, we have*

$$\mathcal{R}(h\circ f;\mathcal{D}_\pi) \leq c\|h\|\sqrt{K\sigma}(\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi))^{\frac{1}{4}} + \|h\|\tau(\sigma,\delta) + \sum_{k=1}^{K} \mathcal{D}_\pi(C_k)\|e_k - h\circ\mu_k(f;\mathcal{D}_\pi)\|, \quad (3)$$

where c is an absolute constant and $\tau(\sigma,\delta)$ is a constant which only depends on (σ,δ). The detail form of τ is shown in the appendix.

The first term in equation 3 is the alignment of f, which is optimized during the contrastive learning pre-training. The second term is determined solely by the (σ,δ) quantity of the data augmentation. Larger σ and smaller δ induce smaller $\tau(\sigma,\delta)$. These two terms together quantify the cluster level of the representation f learned by contrastive learning. The third term is related to the linear layer h on top of f, and will be minimized in downstream training, where each class center is converted to the corresponding ground-truth label. If the regularization term $\mathcal{L}_{\text{reg}}$ is appropriately chosen, the class centers can be distinguished from each other, and the third term can be reduced to 0 by h.

Taking the linear classifier into account, we obtain the following main result.

Theorem 3.2. *Suppose that $\mathcal{A}$ is a (σ,δ)-transformation on $\mathcal{D}$, and π is the augmentation distribution on $\mathcal{A}$. For an encoder f, define*

$$\gamma_{\text{reg}}(f;\mathcal{D},\pi) = 1 - c_1(\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi))^{1/4} - \tau - c_2 \max_{k\neq k'} |\mu_k(f;\mathcal{D}_\pi)^\top \mu_{k'}(f;\mathcal{D}_\pi)|$$

for some positive constants c_1 and c_2. Then for any f such that $\{\mu_k(f;\mathcal{D}_\pi)\}_{k=1}^K$ are linearly independent and $\gamma_{reg}(f;\mathcal{D},\pi) > 0$, we have

$$\mathcal{R}(f;\mathcal{D}_\pi) \leq c\cdot\left(\gamma_{\text{reg}}^{-1}(f;\mathcal{D},\pi)\,(\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi))^{1/4} + \gamma_{\text{reg}}^{-1}(f;\mathcal{D},\pi)\tau(\sigma,\delta)\right), \quad (4)$$

where c is an absolute constant.

Remark 1. *(a) The linear independence of $\{\mu_k(f;\mathcal{D}_\pi)\}_{k=1}^K$ condition is weak, since its complement is a null set in Euclidean space under Lebesgue measure, and have 0 probability under any distribution that is absolutely continuous w.r.t. Lebesgue measure.*

(b) The quantity γ_{reg}^{-1} is essentially the bound of l_2 norm of the linear classifier h which reduces the third term $\sum_{k=1}^K \mathcal{D}_\pi(C_k)\|e_k - h\circ\mu_k(f;\mathcal{D}_\pi)\|$ in equation 3 to 0, i.e., mapping the class centers to their corresponding ground-truth labels. It can be viewed as a measure of "diversity" of the encoder, which encourages the class centers to be orthogonal. The more orthogonal the class centers, the larger the γ_{reg} and the better the bounds. The $\mathcal{L}_{\text{align}}$ term involved in γ_{reg} indicates the phenomenon that better alignment can loose the requirement of diversity. The theory developed in (Huang et al., 2022) shows that popular algorithms such as SimCLR and Barlow Twins can indeed push away the class centers, and the constant γ_{reg} can be ensured to have a constant gap away from 0. Therefore, for simplicity we directly impose this requirement on f.

This theorem shows that contrastive learning on distribution $\mathcal{D}$ with augmentation π is essentially optimizing the supervised loss (ERM) on distribution $\mathcal{D}_\pi$. Therefore, the diversity of π is crucial to

the OOD generalization performance of contrastive. The more diverse π is, we can expect $\mathcal{D}_\pi$ to cover more downstream distributions, and contrastive learning has better OOD performance.

For example, consider the case where original data $\mathcal{D}$ are gray digits, and the downstream data $\mathcal{D}_{\text{red}}$ are obtained by dying them red. Then if we use random coloring as data augmentation and conduct contrastive learning on $\mathcal{D}$, the learned feature will generalize to $\mathcal{D}_{\text{red}}$ well, even though $\mathcal{D}_{\text{red}}$ is far away from $\mathcal{D}$.

4 Out-of-distribution Limitations of Contrastive learning: A Domain Invariance View

Our theoretical results in the last section suggests that data augmentation is crucial to the OOD performance of contrastive learning. However, there exists a fundamental limitation of contrastive learning: *the learned representation is not domain invariant,* which is important for a model to generalize to unseen test datasets (Arjovsky et al., 2019; Peters et al., 2016).

Roughly speaking, domain invariance means that the learned model can extract some intrinsic features, which are invariant across different domains. Therefore, these features can also be applied to new domains. In supervised domain generalization, a common setting is to have a training domain set consisting of multiple source domains. In contrastive learning, naturally, we can also consider $\mathcal{D}_A$ as a *transformation-induced domain* for each data transformation A, and consider $\{\mathcal{D}_A\}_{A\in\mathcal{A}}$ as the training domain set. Since the goal of contrastive learning is also to align different $\mathcal{D}_A$, we naturally expect that the features it learn are domain invariant. However, since the alignment loss L_{align} in contrastive learning is obtained by averaging (i.e., taking expectation), the learned features are not invariant even on $\{\mathcal{D}_A\}_{A\in\mathcal{A}}$. In other words, encoders learned through contrastive learning could behave extremely differently in different $\mathcal{D}_A$.

The following toy model gives an illustrative example.

Proposition 4.1. *Consider a two-dimensional classification problem with data $(X_1, X_2) \sim \mathcal{N}(0, I_2)$. The label Y satisfies $Y = 1(X_1 \geq 0)$, and the data augmentation is to multiply X_2 by standard normal noise, i.e.,*

$$A_\theta(X) = (X_1, \theta \cdot X_2),$$
$$\theta \sim \mathcal{N}(0, 1).$$

The corresponding transformation-induced domain set is $\mathcal{P} = \{\mathcal{D}_c : \mathcal{D}_c = (X_1, c \cdot X_2) \text{ for } c \in \mathbb{R}\}$. We consider the 0-1 loss in equation 2. Then for every $\varepsilon > 0$, there exists representation f and two domains $\mathcal{D}_c$ and $\mathcal{D}_{c'}$ such that

$$\mathcal{L}_{\text{align}}(f; \mathcal{D}, \pi) < \varepsilon,$$

but

$$|\mathcal{R}(f; \mathcal{D}_c) - \mathcal{R}(f; \mathcal{D}_{c'})| \geq \frac{1}{4}$$

This example shows that there exists a representation with arbitrary small contrastive loss, but with very different performance over different transformation-induced domains. The intuition behind this example is that in order to make $\mathcal{L}_{\text{align}}$ small, different domains (induced by different augmentations) need to be aligned in an average rather than uniform sense. Therefore, the representation may still suffer a large alignment loss on some rarely selected augmented domains.

5 ArCL: Augmentation-Robust Contrastive Learning

The above analysis demystifies the OOD generalization ability of contrastive learning on both the pros and cons. On the one hand, the data augmentation used by contrastive learning extends the training dataset, thus enabling generalization to more downstream tasks. On the other hand, contrastive learning is still unable to learn domain-invariant features, even on transformation-induced domains.

To further improve the OOD generalization performance of contrastive learning, we consider how to learn a representation that is invariant across $\{\mathcal{D}_A\}_{A\in\mathcal{A}}$. First of all, we need a definition that mathematically characterize domain invariance, so that we can design algorithms based on it. Here, we borrow the following notion proposed by (Arjovsky et al., 2019).

Definition 5.1. *We say that a data representation f elicits an invariant predictor $h_0 \circ f$ across a domain set $\mathcal{P}$, if there is a classifier h_0 simultaneously optimal for all domains in $\mathcal{P}$, that is, $h_0 \in \arg\min_h \mathcal{R}(h \circ f; \mathcal{D})$ for all $\mathcal{D} \in \mathcal{P}$.*

This definition is equivalent to learning features whose correlations with the target variable are stable, and has been shown to improve OOD generalization both practically and theoretically. Interestingly, by setting $\mathcal{P}$ to be $\{\mathcal{D}_A\}_{A \in \mathcal{A}}$, this definition is well suited to contrastive learning. Indeed, define the *augmentation-robust loss* as

$$L_{\mathrm{AR}}(f; \mathcal{D}) := \mathop{\mathbb{E}}_{X \in \mathcal{D}} \sup_{A, A' \in \mathcal{A}} \|f(A(X)) - f(A'(X))\|^2,$$

we have the following results.

Proposition 5.1. *For any two transformations A and A', linear predictor h and representation f, we have*

$$\sup_{A, A' \in \mathcal{A}} |\mathcal{R}(h \circ f; \mathcal{D}_A) - \mathcal{R}(h \circ f; \mathcal{D}_{A'})| \leq c \cdot \|h\| L_{\mathrm{AR}}(f, \mathcal{D}). \tag{5}$$

Moreover, fix f and let $h_A = \arg\min_h \mathcal{R}(h \circ f, \mathcal{D}_A)$. Then we have

$$|\mathcal{R}(h_A \circ f; \mathcal{D}_{A'}) - \mathcal{R}(h_{A'} \circ f; \mathcal{D}_{A'})| \leq 2c \cdot (\|h_A\| + \|h_{A'}\|) L_{\mathrm{AR}}(f, \mathcal{D}). \tag{6}$$

$L_{\mathrm{AR}}(f; \mathcal{D})$ is an enhanced version of the original alignment loss L_{align}. Compared to L_{align}, L_{AR} replaces the expectation over $\mathcal{A}$ by the supremum, hence $L_{\mathrm{align}}(f; \mathcal{D}, \pi) \leq L_{AR}(f; \mathcal{D})$ for all f and π. If $L_{\mathrm{AR}}(f)$ is small, the above proposition shows that $\mathcal{R}(h \circ f; \mathcal{D}_A)$ varies slightly with A, so that the optimal h for $\mathcal{D}_A$ is close to that for $\mathcal{D}_{A'}$. In other words, representation with smaller L_{AR} tends to *elicit the same linear optimal predictors* across different domains, a property that does not hold for standard alignment loss. Small L_{AR} not only requires f to push the positive pairs closer, but also forces this alignment to be uniform across all transformations in $\mathcal{A}$.

Therefore, we propose to train the encoder using the augmentation-robust loss for better OOD performance, i.e., replace L_{align} by L_{AR} in contrastive loss equation 1. In practical implementation, $\sup_{A_1, A_2 \in \mathcal{A}} \|f(A_1(X)) - f(A_2(X))\|_2^2$ can not be computed exactly, unless $|\mathcal{A}|$ is finite and small. Therefore, we use the following approach to approximate it. For each sample X, we first randomly select m augmentations to obtain a set of augmented samples denoted by $\widehat{\mathcal{A}}_m(X)$. Then, rather than taking average as in standard contrastive learning, we consider use

$$\sup_{A_1, A_2 \in \widehat{\mathcal{A}}_m(X)} \|f(A_1(X)) - f(A_2(X))\|_2^2$$

as the alignment term, which is an approximation of $\sup_{A_1, A_2 \in \mathcal{A}} \|f(A_1(X)) - f(A_2(X))\|_2^2$. The integer m is called the *number of views*, which controls the approximation accuracy. Then, given n unlabeled samples $(X_1, \ldots, X_n)$, the empirical augmentation-robust loss is

$$\widehat{L}_{\mathrm{AR}}(f) := \frac{1}{n} \sum_{i=1}^{n} \sup_{A_1, A_2 \in \widehat{\mathcal{A}}_m(X_i)} \|f(A_1(X_i)) - f(A_2(X_i))\|_2^2. \tag{7}$$

We remark that our proposed method can be applied to any contrastive learning methods whose objective can be formulated by equation 1, such as SimCLR, Moco, Barlow Twins (Zbontar et al., 2021) and so on. We show SimCLR + ArCL as an example in Algorithm 1, and provide MoCo + ArCL in the Appendix. The difference between Algorithm 1 and SimCLR is only the construction of positive pairs. In each epoch t, for every sample X, we first randomly select m transformations denoted by $\widehat{\mathcal{A}}(X) = \{A_1, \ldots, A_m\}$. Then based on current encoder f and projector g, we select two transformations that have worst alignment (minimal inner product) among all pairs in $\widehat{\mathcal{A}}(X)$, and use them to construct the finally used positive pairs of X. The construction of negative pairs and update rule are the same as SimCLR.

5.1 Theoretical Justification

We now give theoretical guarantees of our proposed approximation, i.e., bound the gap between $\widehat{L}_{\mathrm{AR}}(f)$ and $L_{\mathrm{AR}}(f)$. For each representation $f \in \mathcal{F}$, where $\mathcal{F}$ is some hypothesis space, define $z_f(x) := \sup_{A_1, A_2 \in \mathcal{A}} \|f(A_1 x) - f(A_2 x)\|_2$ and $\mathcal{Z}_{\mathcal{F}} := \{z_f : f \in \mathcal{F}\}$. Then we use the

Algorithm 1: SimCLR + ArCL

input : Batch size N, temperature τ, augmentation π, number of views m, epoch T, encoder f, projector g.

for $t = 1, \ldots, T$ **do**
  sample minibatch $\{X_i\}_{i=1}^N$;
  for $i = 1, \ldots, N$ **do**
    draw m augmentations $\widehat{\mathcal{A}} = \{A_1, \ldots, A_m\} \sim \pi$;
    $z_{i,j} = g(f(A_j X_i))$ for $j \in [m]$;
    # select the worst positive samples;
    $s_i^+ = \min_{j,k \in [m]} \{z_{i,j}^\top z_{i,k} / (\|z_{i,j}\| \|z_{i,k}\|)\}$;
    # select the negative samples;
    for $j = 1, \ldots, N$ **do**
      $s_{i,j}^- = z_{i,1}^\top z_{j,1} / (\|z_{i,1}\| \|z_{j,1}\|)$;
      $s_{i,j+N}^- = z_{i,1}^\top z_{j,2} / (\|z_{i,1}\| \|z_{j,2}\|)$;
  compute $L = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_i^+/\tau)}{\sum_{j=1, j \neq i}^{2N} \exp(s_{i,j}^-/\tau)}$;
  update f and g to minimize L;
return f

Rademacher complexity of $\mathcal{Z}$ to define the *alignment complexity* of $\mathcal{F}$ as

$$\widetilde{\mathfrak{R}}_n(\mathcal{F}) := \mathfrak{R}_n(\mathcal{Z}_{\mathcal{F}}) = \mathbb{E}\left[\sup_{z \in \mathcal{Z}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i z(X_i)\right)\right].$$

This quantity is a characteristic of the complexity of $\mathcal{F}$ *in terms of its alignment*. If most of the functions in $\mathcal{F}$ have good alignment under $\mathcal{A}$, then $\widetilde{\mathfrak{R}}_n(\mathcal{F})$ is small even though $\mathfrak{R}_n(\mathcal{F})$ could be large. Using this definition, we have the following result.

Theorem 5.2. *Suppose that $\Theta = B_{d_{\mathcal{A}}}(1)$, i.e., the $d_{\mathcal{A}}$-dimensional unit ball. Suppose that all A in $\mathcal{A}$ are $L_{\mathcal{A}}$-Lipschitz continuous. Let π be a distribution on Θ with density p_π such that $\inf_{\theta \in \Theta} p_\pi(\theta) \geq c_\pi$. Then with probability at least $1 - 2\varepsilon$, for every $f \in \mathcal{F}$,*

$$L_{\mathrm{AR}}(f) \leq \widehat{L}_{\mathrm{AR}}(f) + 2\widetilde{\mathfrak{R}}_n(\mathcal{F}) + \left(\frac{\log(1/\varepsilon)}{n}\right)^{\frac{1}{2}} + L L_A \left(\frac{\log(2n/\varepsilon)}{c_\pi m}\right)^{\frac{1}{d_{\mathcal{A}}}}.$$

The bound in Theorem D.1 consists of two terms $O(n^{-1/2})$ and $O(m^{-1/d_{\mathcal{A}}})$. The first one is due to finite samples, and the second is caused by the finite maximization approximation of the supremum. As m increases, the second term decreases and the bound becomes tighter. Note that we need to assume that the augmentation distribution π behaves uniformly, i.e., $\inf_{\theta \in \Theta} p_\pi(\theta) \geq c$ for some positive constant c, otherwise we can not ensure that the finite maximization is an acceptable approximation.

6 Experiments

We conduct experiments on several OOD settings to show the performance of our proposed methods. To show the adaptability of our approach, we apply ArCL on SimCLR (Chen et al., 2020) and MoCo V2 (He et al., 2020). Following the setting of SimCLR and MoCo V2, we add a projection head after the backbone in the pretraining stage and remove it in the downstream task. For the linear evaluation, we follow the protocol that the encoder is fixed and a linear classifier on the top of the encoder is trained using test set. For the finetuning, we also add a linear classifier on the top of the encoder but we do not fix the encoder.

6.1 Experiments on Modified CIFAR10 and CIFAR100

Setup. Our representations are trained on CIFAR-10 (Krizhevsky, 2009) using SimCLR modified with our proposed ArCL. Different batch sizes $(256, 512)$ and number of views $(m = 4, 6, 8)$ are

Table 1: 5 different augmentations.

	Grayscale	RandomCrop	HorizontalFlip	ColorJitter
Aug 1	✓	–	–	–
Aug 2	–	✓	–	–
Aug 3	–	–	✓	–
Aug 4	–	–	–	✓
Aug 5	✓	–	–	✓

Table 2: Linear evaluation results of pretrained CIFAR10 models on CIFAR10, CIFAR100 and their modified versions.

	Method	Batch Size	Aug 1	Aug 2	Aug 3	Aug 4	Aug 5	Original
CIFAR10	SimCLR	256	83.21	86.93	86.42	86.36	86.13	86.76
	SimCLR + ArCL (views=4)	256	86.77	89.01	88.70	88.68	88.31	88.95
	SimCLR + ArCL (views=6)	256	**87.18**	**89.54**	**88.92**	**88.95**	**88.61**	**89.11**
	SimCLR	512	86.27	88.96	88.56	88.62	88.37	88.81
	SimCLR + ArCL (views=4)	512	88.06	90.48	89.91	89.97	89.59	90.20
	SimCLR + ArCL (views=6)	512	88.94	90.69	90.43	90.24	90.07	90.69
	SimCLR + ArCL (views=8)	512	**88.96**	**90.98**	**90.63**	**90.44**	**90.31**	**90.84**
CIFAR100	SimCLR	256	51.36	51.65	47.55	53.17	52.05	52.75
	SimCLR+ArCL(views=4)	256	52.96	53.76	49.80	55.68	54.19	54.83
	SimCLR+ArCL(views=6)	256	**53.46**	**54.13**	**50.74**	**55.74**	**54.75**	**55.29**
	SimCLR	512	51.53	52.28	48.09	53.45	52.58	53.12
	SimCLR+ArCL(views=4)	512	52.61	53.40	50.16	54.92	53.77	54.20
	SimCLR+ArCL(views=6)	512	**53.77**	54.00	50.57	**56.24**	**55.04**	55.60
	SimCLR+ArCL(views=8)	512	53.21	**54.59**	**50.85**	55.74	54.62	**55.96**

considered. We use ResNet-18 as the encoder and a 2-layer MLP as the projection head. We train the representation with 500 epochs. The temperature is 0.5. Warm-up is not used. We conduct linear evaluation with 100 epochs on different downstream datasets to test the OOD performance of the learned representations.

Datasets. We use CIFAR10 to train the representation. The data augmentation used follows the setting in SimCLR, i.e., a composition of RandomCrop, RandomHorizontalFlip, ColorJitter and Random Grayscale. To evaluate the OOD generalization ability of the learned representation, we use different downstream datasets to train the linear classifier and test its accuracy. Downstream datasets are generated by modifying the original CIFAR10 and CIFAR100 datasets, i.e., applying different data augmentations to them. We choose 5 different augmentations as shown in Table 1. The results are shown in Table 2.

Results. In all settings, our proposed approach improves the OOD performance of SimCLR significantly. As the number of views m increases, the accuracy also increases, which matches our theoretical results. Besides, performance improvements tend to saturate, which indicates that m needs not be very large.

6.2 Experiments on ImageNet Transferring to Small Datasets

Setup. In this part, we will first train an encoder on ImageNet(Deng et al., 2009) using MoCo v2 modified with our proposed ArCL. We select the number of ArCL views from $\{2, 3, 4\}$. Notice that since there exists asymmetric network architecture in MoCo, there is difference between MoCo and MoCo + ArCL(views=2). We use ResNet50 as the backbone and follow all the other architecture settings of MoCo. The memory bank size K is set to 65536 and the batch size is set to 256. For the training epochs, we propose three schemes: 1) 200 epochs training from scratch, 2) 50 epochs training from the model which is pretrained by MoCo for 800 epochs, 3) 100 epochs training from the model which is pretrained by MoCo for 800 epochs. For each setting, we search the best start learning rate in $\{0.03, 0.0075, 0.0015\}$ and the best temperature in $\{0.20, 0.15, 0.12, 0.10, 0.05\}$. We use the SGD optimizer and apply the cosine learning rate decay.

After the model has been pretrained by contrastive learning, we test its OOD generalization ability with linear evaluation and full finetuning on varieties of small datasets, following the settings in (Ericsson et al., 2021). For the target datasets, we adopt FGVC Aircraft (Maji et al., 2013), Caltech-

Table 3: Results comparison on linear evaluation (up) and finetuning (down) of pretrained ImageNet models on popular recognition datasets. Additionally, we provide a supervised baseline for comparison, a standard pretrained ResNet50 available from the PyTorch library. The supervised baseline results are from (Ericsson et al., 2021). Results style: **best** in the same epoch setting.

		Epochs	ImageNet	Aircraft	Caltech101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food	Pets	**Avg**
Linear	MoCo	800	70.68	41.79	87.92	39.31	92.28	74.90	73.88	90.07	68.95	83.30	72.49
	MoCo	800+50	70.64	41.00	87.63	39.01	92.27	75.14	74.31	88.31	68.57	83.69	72.21
	MoCo + ArCL(views=2)	800+50	69.70	44.29	89.79	42.15	93.07	76.70	74.20	90.40	70.94	83.68	73.91
	MoCo + ArCL(views=3)	800+50	69.80	44.57	89.48	42.11	**93.29**	**77.33**	74.63	91.13	71.16	**84.23**	74.21
	MoCo + ArCL(views=4)	800+50	69.80	**44.62**	**89.66**	**42.88**	93.22	76.83	**75.00**	**91.59**	**71.35**	83.99	**74.35**
	MoCo	800+100	70.64	41.65	87.64	39.31	92.12	75.03	73.94	89.53	68.31	83.55	72.34
	MoCo + ArCL(views=2)	800+100	70.26	41.86	89.52	40.21	92.64	75.73	**74.04**	88.97	70.06	84.31	73.04
	MoCo + ArCL(views=3)	800+100	70.92	43.87	89.36	**42.37**	93.30	76.93	72.93	90.50	**71.14**	84.03	73.83
	MoCo + ArCL(views=4)	800+100	69.42	**45.55**	**89.61**	41.91	**93.55**	**77.05**	**74.04**	**91.17**	70.93	**85.48**	**74.37**
	MoCo	200	67.72	40.02	86.59	37.41	90.90	72.43	73.88	87.97	66.97	80.00	70.69
	MoCo + ArCL(views=2)	200	68.04	42.34	87.92	36.45	92.29	71.71	74.68	89.00	67.87	81.42	71.85
	MoCo + ArCL(views=3)	200	68.74	**43.21**	88.26	38.27	92.49	75.02	74.68	**89.61**	68.31	**81.64**	72.39
	MoCo + ArCL(views=4)	200	68.92	41.65	**88.42**	**38.77**	**92.70**	**75.73**	**75.43**	88.95	**68.63**	81.60	**72.43**
	Supervised*		77.20	43.59	90.18	44.92	91.42	73.90	72.23	89.93	69.49	91.45	74.12
Finetune	MoCo	800		83.56	82.54	85.09	95.89	71.81	69.95	95.26	76.81	88.83	83.30
	MoCo	800+50		83.15	84.50	85.90	96.13	72.58	70.16	94.44	79.34	86.12	83.59
	MoCo + ArCL(views=2)	800+50		**86.05**	87.38	**87.28**	96.33	79.39	72.18	95.89	81.36	89.03	86.10
	MoCo + ArCL(views=3)	800+50		84.03	87.64	86.34	**96.88**	80.98	72.87	**96.14**	**81.90**	89.20	86.22
	MoCo + ArCL(views=4)	800+50		84.19	**88.42**	86.67	96.68	**81.17**	**73.09**	95.90	81.70	**89.52**	**86.37**
	MoCo	800+100		83.18	84.50	84.27	96.01	72.14	70.27	95.53	78.23	88.73	83.65
	MoCo + ArCL(views=2)	800+100		84.45	86.84	87.20	96.40	78.40	71.91	95.93	80.54	88.56	85.58
	MoCo + ArCL(views=3)	800+100		**85.94**	86.85	**87.34**	96.36	79.75	71.44	96.00	81.48	88.26	85.94
	MoCo + ArCL(views=4)	800+100		85.65	**88.50**	86.39	**96.91**	**81.29**	**73.35**	**96.17**	**81.82**	**89.30**	**86.60**
	MoCo	200		83.18	82.66	84.47	95.51	72.54	70.43	94.99	77.39	86.12	83.03
	MoCo + ArCL(views=2)	200		81.09	83.93	**86.54**	95.88	76.18	**70.69**	94.44	76.78	86.98	83.61
	MoCo + ArCL(views=3)	200		84.79	85.61	85.39	**96.56**	**78.81**	70.59	**95.84**	**80.71**	87.91	85.13
	MoCo + ArCL(views=4)	200		**84.88**	**86.19**	85.90	96.35	78.62	**70.69**	95.77	80.46	**88.00**	**85.21**
	Supervised*			83.50	91.01	82.61	96.39	82.91	73.30	95.50	84.60	92.42	86.92

101 (Fei-Fei et al., 2004), Stanford Cars (Krause et al., 2013), CIFAR10, CIFAR100, DTD (Cimpoi et al., 2014), Oxford 102 Flowers (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014) and Oxford-IIIT Pets (Parkhi et al., 2012). For linear evaluation, multinomial logistic regression is fit on the extracted features. When finetuning, the models are trained for 5,000 steps using SGD with Nesterov momentum. We calculate the average results of the nine downstream datasets.

Results. We compare our methods with the MoCo baseline. There are about an average rise of 2% in the linear evaluation setting and about an average rise of 3% in the finetuning setting. In particular, we gain a huge improvement on CIFAR100 finetuning, where our proposed methods can outperform the baseline for up to 10%. It is also worthy to mention that our methods can reach the supervised results on nearly all the datasets. These facts can all indicate the effectiveness of our proposed methods.

Besides, the results also show that, under the same epoch setting, the average performance grows as the number of the views increase, which fits our theoretical analysis. One can also notice that there is only a small improvement in the $800 + 100$ epoch setting compared to the $800 + 50$ epoch setting. This implies that our methods can also converge in a rather fast speed, which can also be effective within a few epochs of training.

7 Discussion

In this paper, we theoretically study the OOD generalization ability of contrastive learning. Our results suggest that the data augmentation plays an important role in its OOD performance. We also find a limitation of contrastive learning: it cannot learn domain invariant representations. Based on these theoretical results, we propose the Augmentation-robust Contrastive Learning (ArCL), a module that can be applied to various popular contrastive learning methods. It forces the representation to align positive samples uniformly, in order to improve its OOD performance. Our experiments on CIFAR10 and ImageNet verify our theoretical results and the superiority of our method.

REFERENCES

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

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Proc. of NeurIPS*, 2011.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. of CVPR*, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 2021.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5414–5423, June 2021.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004. doi: 10.1109/CVPR.2004.383.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Proc. of NeurIPS*, 2020.

Jeff Z HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. of CVPR*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning, 2022.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Proc. of NeurIPS*, 2020.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proc. of ICML*, 2018.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Proc. of NeurIPS*, 2020.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

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 *Proc. of ICML*, 2021.

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proc. of AAAI*, 2020.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *arXiv preprint arXiv:2011.01681*, 2020.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Proc. of ICML*, 2021.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL `https://arxiv.org/abs/1306.5151`.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proc. of ICML*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2016.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Proc. of ICLR*, 2019.

Kendrick Shen, Robbie Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2204.00570*, 2022.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kuang. Stable learning via sample reweighting. In *Proc. of AAAI*, 2020.

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.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proc. of ECCV*, 2020.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proc. of ICML*, 2021.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *arXiv preprint arXiv:2106.04619*, 2021.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021a.

Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Proc. of NeurIPS*, 2021b.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. of ICML*, 2020.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *arXiv preprint arXiv:2106.04496*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proc. of ICML*, 2021.

A MoCo + ArCL

Algorithm 2: MoCo + ArCL

input : Batch size N, temperature τ, memory bank M, augmentation π, number of views m, epoch T, encoders f_q, f_k, projector g.

for $t = 1, \ldots, T$ **do**
  sample minibatch $\{X_i\}_{i=1}^N$;
  for $i = 1, \ldots, N$ **do**
    draw m augmentations $\widehat{\mathcal{A}} = \{A_1, \ldots, A_m\} \sim \pi$;
    $z_{i,j}^1 = g(f_q(A_j X_i))$ for $j \in [m]$;
    $z_{i,j}^2 = g(f_k(A_j X_i))$ for $j \in [m]$;
    # select the worst positive samples;
    $s_i^+ = \min_{j,l \in [m], j \neq l} \{z_{i,j}^{1\top} z_{i,l}^2 / (\|z_{i,j}^1\| \|z_{i,l}^2\|)\}$;
    # select the negative samples;
    for $m \in M$ **do**
      $s_{i,m}^- = z_{i,1}^\top m / (\|z_{i,1}\| \|m\|)$;
  compute $L = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_i^+/\tau)}{\sum_{m \in M} \exp(s_{i,m}^-/\tau)}$;
  update f and g to minimize L;
  update M;
return f

B Proofs in Section 3

Proof of Lemma 3.1. For simplicity we omit the notations $\mathcal{D}$ and ϑ in this proof, and use x_1 to denote the positive sample, i.e. the augmented data, use p_k to denote $\mathcal{D}_\vartheta(C_k)$ and use μ_k to denote $\mu_k(f; \mathcal{D}_\vartheta)$. From Theorem 2 and Lemma B.1 in (Huang et al., 2022), we have

$$\mathop{\mathbb{E}}_{x \in C_k} \|f(x_1) - \mu_k\| \leq c \sqrt{\frac{1}{p_k}} L_{pos}^{\frac{1}{4}}(f) + \tau(\sigma, \delta) \tag{8}$$

for some constant c, where

$$\tau(\sigma, \delta) := 4 \left(1 - \sigma \left(1 - \frac{L\delta}{4}\right)\right).$$

Note that τ is decreasing with σ and increasing with δ. Then we have

$$\begin{aligned}
c \left(\sum_{k=1}^K \sqrt{p_k}\right) L_{\text{pos}}^{\frac{1}{4}}(f) + \tau(\sigma, \delta) &\geq \sum_{k=1}^K p_k \mathop{\mathbb{E}}_{x \in C_k} \|f(x_1) - \mu_k\| \\
&\geq \sum_{k=1}^K \frac{p_k}{\|h\|} \mathop{\mathbb{E}}_{x \in C_k} \mathop{\mathbb{E}}_{x_1 \in Ax} \|h \circ f(x_1) - h \circ \mu_k\| \\
&\geq \sum_{k=1}^K \frac{p_k}{\|h\|} \mathop{\mathbb{E}}_{x \in C_k} \mathop{\mathbb{E}}_{x_1 \in Ax} \|h \circ f(x_1) - e_k\| - \frac{1}{\|h\|} \sum_{k=1}^K p_k \|e_k - h \circ \mu_k\| \\
&= \frac{1}{\|h\|} R(h \circ f) - \frac{1}{\|h\|} \sum_{k=1}^K p_k \|e_k - h \circ \mu_k\|
\end{aligned} \tag{9}$$

for all linear layer $h \in \mathbb{R}^{K \times d_1}$. Therefore, we obtain

$$\begin{aligned} R(h \circ f) &\leq c\|h\| L_{\text{pos}}^{\frac{1}{4}} \sum_{k=1}^{K} \sqrt{p_k} + \|h\| \tau(\sigma, \delta) + \sum_{k=1}^{K} p_k \|e_k - h \circ \mu_k\| \\ &\leq c\|h\| \sqrt{K} L_{\text{pos}}^{\frac{1}{4}} + \|h\| \tau(\sigma, \delta) + \sum_{k=1}^{K} p_k \|e_k - h \circ \mu_k\|. \end{aligned} \tag{10}$$

□

Proof of Theorem 3.2. By the triangle inequality and equation 8, we have

$$\|\mu_k\| \geq 1 - \mathop{\mathbb{E}}_{x \in C_k} \|f(x_1) - \mu_k\| \geq 1 - C\sqrt{\frac{1}{p_k}} L_{pos}^{\frac{1}{4}}(f) - \tau(\sigma, \delta).$$

Therefore,

$$\min_k \|\mu_k\|_2 \geq 1 - C\sqrt{\frac{1}{p_0}} L_{pos}^{\frac{1}{4}}(f) - \tau(\sigma, \delta).$$

Define $U = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^{d_1 \times K}$ and let $h_0 = U^+$ be the Moore–Penrose inverse of U such that $h_0 U = I_K$. Then we have

$$\|h_0\| = \sup_{x \in \mathbb{R}^{d_1}} \frac{\|x\|}{\|Ux\|}.$$

For every $x = (x_1, \ldots, x_{d_1}) \in \mathbb{R}^{d_1}$ with $\|x\|_2 = 1$, we have

$$\begin{aligned} \|Ux\|_2^2 &= \sum_{i=1}^{d_1} x_i^2 \|\mu_i\|^2 + \sum_{i \neq j} x_i x_j \mu_i^\top \mu_j \\ &\geq \min_i \|\mu_i\|_2^2 - \max_{i,j} |\mu_i^\top \mu_j| \sum_{i \neq j} |x_i x_j| \\ &= \min_i \|\mu_i\|_2^2 - \max_{i,j} |\mu_i^\top \mu_j| \left(\left(\sum_i |x_i| \right)^2 - 1 \right) \\ &\geq \min_i \|\mu_i\|_2^2 - (K-1) \max_{i,j} |\mu_i^\top \mu_j| \\ &\geq 1 - c\sqrt{\frac{1}{p_0}} L_{\text{pos}}^{\frac{1}{4}}(f) - \tau(\sigma, \delta) - K \max_{i,j} |\mu_i^\top \mu_j| \\ &=: \frac{1}{\eta_{\text{reg}}(f; \mathcal{D}, \vartheta)}. \end{aligned}$$

Therefore, we have

$$\|h_0\| \leq \eta_{\text{reg}}(f; \mathcal{D}, \vartheta).$$

Then, by Lemma 3.1, we complete the proof by

$$R(f) = \min_h R(h \circ f) \leq R(h_0 \circ f) \leq c\gamma_{reg}(f; \mathcal{D}, \vartheta) L_{\text{pos}}^{\frac{1}{4}}(f; \mathcal{D}, \vartheta) + \gamma_{reg}(f; \mathcal{D}, \vartheta)\tau(\sigma, \delta).$$

□

C Proof in Section 4

Proof of Proposition 4.1. For any $\varepsilon > 0$, let $t = \sqrt{\varepsilon}/2$ and $f(x_1, x_2) = x_1 + tx_2$. Then, the alignment loss of f satisfies

$$\mathcal{L}_{\text{align}}(f; \mathcal{D}, \pi) = t^2 \, \mathbb{E}\, X_2^2 \mathop{\mathbb{E}}_{(\theta_1, \theta_2) \sim \mathcal{N}(0,1)^2} (\theta_1 - \theta_2)^2 = 2t^2 < \varepsilon.$$

Let $c = 0$ and $c' = 1/t$. Then obviously

$$\mathcal{R}(f; \mathcal{D}_c) = 0,$$

but

$$\mathcal{R}(f; \mathcal{D}_{c'}) = P(X_1 < 0, X_1 + X_2 \geq 0) + P(X_1 \geq 0, X_1 + X_2 \leq 0) = \frac{1}{4}.$$

□

D PROOFS IN SECTION 5

Proof of Proposition 5.1. For any f and h with $\|h\| \leq c_h$, we have

$$\begin{aligned}
R(h \circ f; \mathcal{D}_A) - R(h \circ f; \mathcal{D}_{A'}) &= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \left(|h \circ f(AX) - Y|^2 - |h \circ f(A'X) - Y|^2\right) \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} (h \circ f(Ax) - h \circ f(A'x))((h \circ f(Ax) + h \circ f(A'x)) + 2y) \\
&\leq c \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \|h \circ f(Ax) - h \circ f(A'x)\| \\
&\leq c\|h\| \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \|f(Ax) - f(A'x)\| \\
&\leq c\|h\| L_{AR}(f)
\end{aligned}$$

for some constant c which depends on c_h. □

proof of Theorem D.1. We first fix the sample $(x_1, \ldots, x_n)$, and consider the randomness of $\mathcal{A}_m(x_i)$. Define $A_1^*(x_i)$ and $A_2^*(x_i)$ as

$$A_1^*(x_i), A_2^*(x_i) = \arg \sup_{A_1, A_2 \in \mathcal{A}} \|f(A_1 x_i) - f(A_2 x_i)\|_2^2.$$

Note that their order doesn't matter. For every $\varepsilon > 0$, we have

$$\begin{aligned}
&P\left\{\frac{1}{n}\sum_{i=1}^n \sup_{A_1,A_2\in\mathcal{A}} \|f(A_1x_i) - f(A_2x_i)\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \sup_{A_1,A_2\in\mathcal{A}_m(x_i)} \|f(A_1x_i) - f(A_2x_i)\|_2^2 + L^2 L_{\mathcal{A}}^2 \left(\frac{\log(2n/\varepsilon)}{c_\vartheta m}\right)^{\frac{2}{d}}\right\} \\
\geq& P\left\{\max_{i\in[n]}\left\{\inf_{A\in\mathcal{A}_m(x_i)} \|A - A_1^*(x_i)\|_2, \inf_{A\in\mathcal{A}_m(x_i)} \|A - A_2^*(x_i)\|_2\right\} \leq \left(\frac{\log(2n/\varepsilon)}{c_\vartheta m}\right)^{\frac{1}{d}}\right\} \\
\geq& 1 - \sum_{i=1}^n P\left\{\inf_{A\in\mathcal{A}_m(x_i)} \|A - A_1^*(x_i)\|_2 \leq \left(\frac{\log(2n/\varepsilon)}{c_\vartheta m}\right)^{\frac{1}{d}}\right\} - \sum_{i=1}^n P\left\{\inf_{A\in\mathcal{A}_m(x_i)} \|A - A_2^*(x_i)\|_2 \leq \left(\frac{\log(2n/\varepsilon)}{c_\vartheta m}\right)^{\frac{1}{d}}\right\} \\
\geq& 1 - 2n\left(1 - \frac{c_\vartheta \log(2n/\varepsilon)}{c_\vartheta m}\right)^m \\
\geq& 1 - \varepsilon.
\end{aligned}$$

Therefore, with probability at least $1 - \varepsilon$, we have

$$\frac{1}{n}\sum_{i=1}^n \sup_{A_1,A_2\in\mathcal{A}} \|f(A_1x_i) - f(A_2x_i)\|_2 \leq \widehat{L}_{AR}(f) + L_f^2 L_A^2 \left(\frac{\log(2n/\varepsilon)}{m}\right)^{\frac{2}{d}}.$$

Besides, for every f we have

$$L_{AR}(f) \leq \frac{1}{n}\sum_{i=1}^n \sup_{A_1,A_2\in\mathcal{A}} \|f(A_1x_i) - f(A_2x_i)\|_2^2 + 2\mathfrak{R}_n(\mathcal{F}) + \left(\frac{\log(1/\varepsilon)}{n}\right)^{\frac{1}{2}}$$

with probability at least $1 - \varepsilon$. Then we finally obtain that with probability at least $1 - 2\varepsilon$, for every $f \in \mathcal{F}$,

$$L_{AR}(f) \leq \widehat{L}_{AR}(f) + 2\widetilde{\mathfrak{R}}_n(\mathcal{F}) + \left(\frac{\log(1/\varepsilon)}{n}\right)^{\frac{1}{2}} + L_f^2 L_A^2 \left(\frac{\log(2n/\varepsilon)}{c_\vartheta m}\right)^{\frac{2}{d}}.$$

□

D.1 THEORETICAL JUSTIFICATION OF THE EMPIRICAL AUGMENTATION-ROBUST LOSS

We now give theoretical guarantees of our proposed approximation, i.e., bound the gap between $\widehat{L}_{\mathrm{AR}}(f)$ and $L_{\mathrm{AR}}(f)$. For each representation $f \in \mathcal{F}$, where $\mathcal{F}$ is some hypothesis space, define $z_f(x) := \sup_{A_1,A_2\in\mathcal{A}} \|f(A_1x) - f(A_2x)\|_2$ and $\mathcal{Z}_{\mathcal{F}} := \{z_f : f \in \mathcal{F}\}$. Then we use the

Rademacher complexity of $\mathcal{Z}$ to define the *alignment complexity* of $\mathcal{F}$ as

$$\widetilde{\mathfrak{R}}_n(\mathcal{F}) := \mathfrak{R}_n(\mathcal{Z}_{\mathcal{F}}) = \mathbb{E}\left[\sup_{z\in\mathcal{Z}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i z(X_i)\right)\right].$$

This quantity is a characteristic of the complexity of $\mathcal{F}$ *in terms of its alignment*. If most of the functions in $\mathcal{F}$ have good alignment under $\mathcal{A}$, then $\widetilde{\mathfrak{R}}_n(\mathcal{F})$ is small even though $\mathfrak{R}_n(\mathcal{F})$ could be large. Using this definition, we have the following result.

Theorem D.1. *Suppose that $\Theta = B_{d_{\mathcal{A}}}(1)$, i.e., the $d_{\mathcal{A}}$-dimensional unit ball. Suppose that all A in $\mathcal{A}$ are $L_{\mathcal{A}}$-Lipschitz continuous. Let π be a distribution on Θ with density p_π such that $\inf_{\theta\in\Theta} p_\pi(\theta) \geq c_\pi$. Then with probability at least $1-2\varepsilon$, for every $f \in \mathcal{F}$,*

$$L_{\mathrm{AR}}(f) \leq \widehat{L}_{\mathrm{AR}}(f) + 2\widetilde{\mathfrak{R}}_n(\mathcal{F}) + \left(\frac{\log(1/\varepsilon)}{n}\right)^{\frac{1}{2}} + LL_A\left(\frac{\log(2n/\varepsilon)}{c_\pi m}\right)^{\frac{1}{d_{\mathcal{A}}}}.$$

The bound in Theorem D.1 consists of two terms $O(n^{-1/2})$ and $O(m^{-1/d_{\mathcal{A}}})$. The first one is due to finite samples, and the second is caused by the finite maximization approximation of the supremum. As m increases, the second term decreases and the bound becomes tighter. Note that we need to assume that the augmentation distribution π behaves uniformly, i.e., $\inf_{\theta\in\Theta} p_\pi(\theta) \geq c$ for some positive constant c, otherwise we can not ensure that the finite maximization is an acceptable approximation.