Domain Generalization through the Lens of Angular Invariance

Yujie Jin[1] * , **Xu Chu**[1,2] * , **Yasha Wang**[1] † and **Wenwu Zhu**[2] †

[1] Peking University, Beijing, China
[2] Tsinghua University, Beijing, China

{jyj17pku,wangyasha}@pku.edu.cn, {chu_xu,wwzhu}@tsinghua.edu.cn

Abstract

Domain generalization (DG) aims at generalizing a classifier trained on multiple source domains to an unseen target domain with domain shift. A common pervasive theme in existing DG literature is domain-invariant representation learning with various invariance assumptions. However, prior works restrict themselves to an impractical assumption for real-world challenges: If a mapping induced by a deep neural network (DNN) could align the source domains well, then such a mapping aligns a target domain as well. In this paper, we simply take DNNs as feature extractors to relax the requirement of distribution alignment. Specifically, we put forward a novel angular invariance and the accompanied norm shift assumption. Based on the proposed term of invariance, we propose a novel deep DG method dubbed Angular Invariance Domain Generalization Network (AIDGN). The optimization objective of AIDGN is developed with a von-Mises Fisher (vMF) mixture model. Extensive experiments on multiple DG benchmark datasets validate the effectiveness of the proposed AIDGN method.

1 Introduction

Over the past few years, supervised deep learning has achieved remarkable success on many challenging visual tasks [Krizhevsky *et al.*, 2012; Long *et al.*, 2015; He *et al.*, 2016]. An underlying assumption of the popular supervised DL methods is the identically distributed condition, namely, the generating functions of training data and testing data are identical. We say a *domain shift* exists between the training data (*source domain*) and the testing data (*target domain*) if the identical condition is violated. When there is a domain shift, the favored empirical risk minimization (ERM) learning [Vapnik, 1999] would be ill-posed, since the empirical risk over the training data is not guaranteed to converge to the risk of the testing data asymptotically.

Domain generalization (DG) aims at generalizing the model trained on multiple source domains to perform well on an un-

---

* The first two authors contributed to this work equally.
† Corresponding authors.

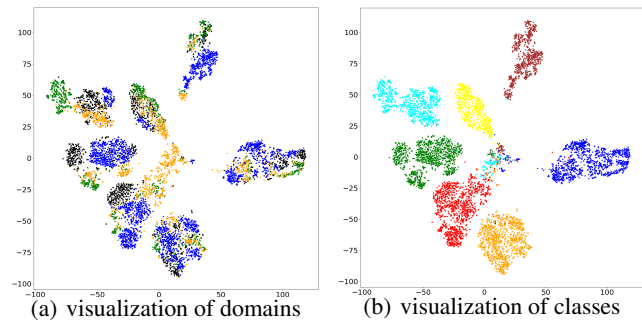|	
(a) visualization of domains	(b) visualization of classes

Figure 1: Feature visualization for a model trained with ERM on the PACS dataset: (a) different colors indicate different domains, source domains include cartoon (black), photo (green) and sketch (blue) while the target domain is art-painting (orange); (b) different colors represent different classes. *Best viewed in color (Zoom in for details).*

seen target domain with domain shift [Blanchard *et al.*, 2011]. The inductive setting of DG assumes no target data is available during training, differentiating DG from the transductive domain adaptation methodss [Ben-David *et al.*, 2007], thus making DG more practical and challenging.

Intuitively, in order to carry out a successful knowledge transfer from "seen" source domains to an "unseen" target domain, there have to be some underlying similarities among these domains. From a theoretical standpoint, invariance among the distributions of domains should be investigated. To this end, a predominant stream in DG is domain-invariant representation learning, with various invariance and shift assumptions such as *covariate shift* assumption [Li *et al.*, 2018b], *conditional shift* assumption [Li *et al.*, 2018c], and *label shift* assumption [Liu *et al.*, 2021]. However, prior works overemphasize the importance of joint distribution alignment under an impractical assumption: an injective mapping (implies a tendency of losing class discriminative information) aligning the source joint distributions on the induced space could align the target joint distribution as well. An easy counter-example is a constant mapping that aligns any distributions on the induced space. Recently, theoretical analysis has revealed a fundamental trade-off between achieving well-alignment and low joint error [Zhao *et al.*, 2019] on various domains. Empirically, a study [Gulrajani and Lopez-Paz, 2021] observed

limited performance gain of those invariant learning methods over ERM under a fair evaluation protocol, demonstrating the difficulty of balancing alignment and generalization.

In this paper, we take a step back from pursuing domain alignment. We model the relative difference between the target domain and each source domain instead. Specifically, we put forward a novel paradigm of domain shift assumption: the **angular invariance** and **norm shift** assumption. The proposed assumption says that under the polar reparameterization [Blumenson, 1960], the relative difference between the DNN push forward measures is captured by the norm parameters and invariant to the angular parameters. The insight of angular invariance and norm shift is inspired by the acknowledged fact that the internal layers in DNNs capture high-level semantic concepts (e.g., eye, tail) [Zeiler and Fergus, 2014], which are connected to category-related discriminative features. The angular parameters capture the correlations between the high-level semantic concepts, while the norm parameter captures the magnitude of the high-level semantic concepts. In the practice of DG, the DNN feature mapping pre-trained on ImageNet is fine-tuned on the source domains. Therefore the semantic concepts memorized by the internal layers are biased to the source domains, and leading to higher-level of neuron activations. Hence we expect a difference of norm distribution of latent representations between a source domain and a target domain. Meanwhile the correlations between high-level concepts in a fixed category are relatively stable. Thus we expect invariant angular distributions across different domains. We do t-SNE feature visualization on the PACS dataset for an ERM-trained model to motivate and substantiate our assumption. Fig 1(a) shows that the norm distribution of the target domain (orange) significantly differs from that of source domains, while the distributions over angular coordinates are homogeneous. Fig 1(b) shows that the learned class clusters are separated well by the angular parameters.

Apart from the novel angular invariance and norm shift assumption, our methodological contribution is manifested by a novel deep DG algorithm called **A**ngular **I**nvariance **D**omain **G**eneralization **N**etwork (**AIDGN**). The designing principle of the AIDGN method is a minimal level of modification of ERM learning under modest intensity distributional assumptions, such as assuming the distribution families of maximum entropy. Concretely, (1) We show that the angular invariance enables us to compute the marginals over the norm coordinate to compare probability density functions of the target distribution and each source distribution in the latent space. Moreover, we compute the relative density ratio analytically based on the maximum entropy principle [Jaynes, 1957]. (2) Within a von-Mises Fisher (vMF) mixture model [Gopal and Yang, 2014], we connect the target posterior with the density of each mixture component, re-weighted by the relative density ratio mentioned above and the label densities. (3) We derive a practical AIDGN loss from the target posterior. The deduction adopts the maximum entropy principle for label densities and solves a constrained optimization problem.

We conduct extensive experiments on multiple DG benchmarks to validate the effectiveness of the proposed method and demonstrate that it achieves superior performance over the existing baselines. Moreover, we show that AIDGN effec-

tively balances the intra-class compactness and the inter-class separation, and thus reduces the uncertainty of predictions.

2 Related Work

A common pervasive theme in DG literature is domain-invariant representation learning, which is based on the idea of aligning feature distributions among different source domains, with the hope that the learned invariance can be generalized to target domains. For instance, [Li *et al.*, 2018b] achieved distribution alignment in the latent space of an autoencoder by using adversarial learning and the maximum mean discrepancy criteria. [Li *et al.*, 2018c] matched conditional feature distributions across domains, enabling alignment of multimodal distributions for all class labels. [Liu *et al.*, 2021] exploited both the conditional and label shifts, and proposed a Bayesian variational inference framework with posterior alignment to reduce both the shifts simultaneously. However, existing works overemphasize the importance of joint distribution alignment which might hurt class discriminative information. Different from them, we propose a novel angular invariance as well as the accompanied norm shift assumption, and develop a learning framework based on the proposed term of invariance.

Meta-learning was introduced into the DG community by [Li *et al.*, 2018a] and has drawn increasing attention. The main idea is to divide the source domains into meta-train-domains and meta-test-domain to simulate domain shift, and regulate the model trained on meta-train-domains to perform well on meta-test-domain. Data augmentation has also been exploited for DG, which augments the source data to increase the diversity of training data distribution. For instance, [Wang *et al.*, 2020b] employed the mixup [Zhang *et al.*, 2018] technique across multiple domains and trained model on the augmented heterogeneous mixup distribution, which implicitly enhanced invariance to domain shifts.

Different from the above DG methods which focus on training phase, test-time adaptation is a class of methods focusing on test phase, i.e., adjusting the model using online unlabeled data and correcting its prediction by itself during test time. [Wang *et al.*, 2020a] proposed fully test-time adaptation, which modulates the BN parameters by minimizing the prediction entropy using stochastic gradient descent. [Iwasawa and Matsuo, 2021] proposed a test-time classifier adjustment module for DG, which updates pseudo-prototypes for each class using online unlabeled data augmented by the base classifier trained on the source domains. We empirically show that AIDGN can effectively make the decision boundaries of all categories separate from each other and reduce the uncertainty of predictions, so that the existing test-time adaptation methods based on entropy minimization is not necessary.

We also show that our proposed AIDGN theoretically justifies and generalizes the recent proposed MAG loss for face recognition [Meng *et al.*, 2021].

3 Methodology

In this section, we first formulate the DG problem. Secondly, we explain the proposed angular invariance and norm shift assumption. Lastly, we introduce our angular invariance domain

generalization network (AIDGN). (*Proofs for this section can be found in Appendix A of the supplementary material.*)

3.1 Problem Formulation

Give N source domains $\{\mathcal{P}^d_{\mathcal{X} \times \mathcal{Y}}\}^N_{d=1}$ subject to $\mathcal{P}^d_{\mathcal{X} \times \mathcal{Y}} \neq \mathcal{P}^{d'}_{\mathcal{X} \times \mathcal{Y}}$ for $\{d, d'\} \subset \{1, 2, \dots, N\}$, and a target domain $\mathcal{P}^t_{\mathcal{X} \times \mathcal{Y}}$ on the input-output space $\mathcal{X} \times \mathcal{Y}$. DG tasks assume $\mathcal{P}^d_{\mathcal{X} \times \mathcal{Y}} \neq \mathcal{P}^t_{\mathcal{X} \times \mathcal{Y}}$ for $d = 1, \dots, N$ and focus on C-class single label classification tasks. Let $\mathcal{H} = \{h_\theta | \theta \in \Theta\}$ be a hypothesis space parametrized by $\theta \in \Theta$. For $d = 1, \dots, N$, there are n_d independently identically distributed instances $\{(\mathbf{x}^d_i, y^d_i)\}^{n_d}_{i=1}$ sampled from the d-th source domain $\mathcal{P}^d_{\mathcal{X} \times \mathcal{Y}}$. The goal of DG is to output a hypothesis $\hat{h} \in \mathcal{H}$ such that the target risk is minimized for a given loss $\ell(h(\cdot), \cdot)$, i.e.,

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{P}^t_{\mathcal{X} \times \mathcal{Y}}} [\ell(h(\mathbf{X}), Y)]. \quad (1)$$

3.2 Angular Invariance and Norm Shift

Celebrated for capturing empirical universal visual features, convolutional neural networks (CNNs) pre-trained on the ImageNet dataset [Deng *et al.*, 2009] have been adopted by a wide range of visual tasks. To take full advantage of a pre-trained CNN π, we regard π as a feature extractor from the original input space $\mathcal{X}$ to a latent representation space $\mathcal{Z}$. Then a hypothesis h comprises a feature extractor π and a classifier f, i.e., $h = f \circ \pi$.

Studies have shown that each dimension of a CNN π output is capturing some abstract concepts (e.g., eye, tail) [Zeiler and Fergus, 2014]. Considering the relationship among concepts of the same class objects in the real-world is stable, the angular invariance and norm shift assumption says that the π-mapped feature of different domains are invariant in the angular coordinates, but varies in the norm coordinate. For simplicity, we introduce a random variable D indexing the d-th source domain if $D = d$. The proposed assumption states as follows.

Assumption 1 (angular invariance). *Suppose the marginal distributions $\{\mathcal{P}^d_{\mathcal{X}}\}^N_{d=1} \cup \{\mathcal{P}^t_{\mathcal{X}}\}$ on the input space $\mathcal{X}$ are continuous. Let $\pi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^n$ be a feature extraction mapping such that the π-push forward probability density funcitons (p.d.f.s) $\{p^d(\mathbf{z})\}^N_{d=1} \cup \{p^t(\mathbf{z})\}$ exist in the latent space $\mathcal{Z}$. Let $(r, \phi_1, \dots, \phi_{n-1}) = g(z_1, \dots, z_n)$ be the polar reparametrization [Blumenson, 1960] of the Cartesian coordinates $\mathbf{z} = (z_1, \dots, z_n)$. The angular invariance assumption for DG is quantified by the equations: Let $\phi = (\phi_1, \dots, \phi_{n-1})$, for $d = 1, \dots, N$,*

$$p(\phi|Y, D = d) = p^t(\phi|Y) \quad (2)$$

The polar reparametrization $g(\cdot)$ is bijective and $p(r, \phi|Y, D) = p(\phi|Y, D)p(r|\phi, Y, D)$, therefore the difference between the target conditional p.d.f. (c.p.d.f.) $p^t(\mathbf{z}|Y)$ and the d-th source c.p.d.f. $p(\mathbf{z}|Y, d)$ is captured by the difference between the norm c.p.d.f.s $p^t(r|\phi, Y)$ and $p(r|\phi, d, Y)$.

Theorem 1. *Suppose $support(p^t(\mathbf{z})) \subset support(p(\mathbf{z}|D))$. If the angular invariance assumption 1 holds, then for $d = 1, \dots, N$, $p^t(\mathbf{z}|Y)/p(\mathbf{z}|D = d, Y)$ exists and satisfies*

$$\frac{p^t(\mathbf{z}|Y)}{p(\mathbf{z}|d, Y)} = \frac{p^t(r|\phi, Y)}{p(r|\phi, d, Y)} \triangleq w(r|\phi, d, y). \quad (3)$$

The theorem 1 says that under the angular invariance assumption, we may reduce the degrees of freedom of comparing target and source c.p.d.f.s from n to 1. However, the aporia of DG is that no target instances could be observed during training. Thus an additional assumption is essential to overcome the zero-sample dilemma. Following the maximum entropy principle [Jaynes, 1957], we adopt the following distributional assumptions on the conditional target and source norms.

Assumption 2 (maximum entropy norm distribution). *Conditioned on $Y = y$ and ϕ, (I) The target norm in space $\mathcal{Z}$ follows a continuous uniform distribution[1] $Uni[\alpha_{y,\phi}, \beta_{y,\phi}]$ with $\delta_{y,\phi} = \beta_{y,\phi} - \alpha_{y,\phi} > 0$, i.e., $p^t(r|y, \phi; \alpha_{y,\phi}, \beta_{y,\phi}) = 1/\delta_{y,\phi}$. (II) The d-th source domain norm in space $\mathcal{Z}$ follows an exponential distribution[2] $Exp[1/\mu_{d,y,\phi}], \mu_{d,y,\phi} > 0$, i.e., $p(r|d, y; \mu_{d,y,\phi}) = 1/\mu_{d,y,\phi} exp(r/\mu_{d,y,\phi}).$*

With the angular invariance and the maximum entropy assumption, we can compare $p^t(\mathbf{z}|y)$ and $p(\mathbf{z}|d, y)$ analytically.

Corollary 1. *When assumption 1 and assumption 2 hold,*

$$w(r|\phi, d, y) = \frac{\mu_{d,y} exp(\frac{r}{\mu_{d,y,\phi}})}{\delta_{y,\phi}} \approx \frac{\mu_{d,y,\phi} + r}{\delta_{y,\phi}}. \quad (4)$$

Recalling that DG aims to learn classifiers, next we consider the behavior when Y varies, i.e., $p^t(y|\mathbf{z})$ and $p(y|\mathbf{z}, D)$.

3.3 The AIDGN Method

Before formally introducing the proposed AIDGN method, We discuss the motivation of adopting the von-Mises Fisher (vMF) mixture model. Specifically, we inspect $p(\mathbf{z}|Y)$ and $p(\phi|Y)$, where $\phi = (\phi_1, \dots, \phi_{n-1})$ is the angular coordinates after a polar reparameterization of $\mathbf{z}$. By the law of total probability, the source c.p.d.f. $p^s(\mathbf{z}|Y)$ decomposes as

$$p^s(\mathbf{z}|Y) = \sum_{d=1}^N p(\mathbf{z}|d, Y)p(d|Y) \propto \sum_{d=1}^N p(r, \phi|d, Y)p(d|Y) \quad (5)$$

When the angular invariance and norm shift assumption 1 holds, the factors $p(\mathbf{z}|d, Y)$ and $p(d|Y)$ might varies w.r.t. the domain index d. Therefore modeling $p^s(\mathbf{z}|Y)$ urges the modeling of $p(\mathbf{z}|d, Y)$ w.r.t. $d = 1, \dots, N$. In sharp contrast, the angular invariance guarantees the modeling of the source c.p.d.f. $p^s(\phi|Y)$ is as easy as any $p(\phi|Y, d)$ for $d = 1, \dots, N$. By eq. (2), $p^s(\phi|Y) = \sum_{d=1}^N p(\phi|Y, d)p(d|Y) = p(\phi|Y, d)\sum_{d=1}^N p(d|Y) = p(\phi|Y, d)$. Therefore, the much simpler assumption choice is a model related to $p^s(\phi)$. Notice that the angular coordinates of the latent representation $\mathbf{z}$ are invariant to the L_2 normalization $\mathcal{G}(\mathbf{z})$, i.e., $\mathcal{G} \circ g : (r, \phi) \mapsto (1, \phi)$, where g is the polar reparameterization and $\mathcal{G}(\mathbf{z})$ is

$$\mathcal{G}(\mathbf{z}) = \mathbf{z}/\sqrt{z_1^2 + z_2^2 + \dots + z_n^2}. \quad (6)$$

The formulation of the proposed AIDGN begins with the vMF mixture assumption on the L_2 normalized $\mathcal{G}(\mathbf{Z})$.

---

[1]The uniform distribution is the maximum (differential) entropy distribution for a continuous random variable with a fixed range.

[2]The exponential distribution is the maximum (differential) entropy distribution with positive support and a fixed expectation.

Assumption 3 (von-Mises Fisher mixture). *Suppose that the assumption 1 is satisfied, let $\mathbf{Z}^* \triangleq \mathcal{G}(\mathbf{Z})$ be the L_2 normalization of the latent representation $\mathbf{Z} = \pi(\mathbf{X})$. In the C-category DG classification setting, $\mathbf{Z}^*$ is assumed to follow a von-Mises Fisher mixture distribution,*

$$p(z^*) = \sum_{y=1}^{C} p(Y = y)p(z^*|y) = \sum_{y=1}^{C} p(y)\mathcal{V}(z^*; \boldsymbol{w}_y, \kappa), \quad (7)$$

such that the posterior p.d.f. $p(y|z, D)$ is invariant to the L_2 normalized posterior $p(y|z^*)$

$$p(y|z, D) = p(y|z^*). \quad (8)$$

The y-th component of the mixture $p(z^*|y) = \mathcal{V}(z^*; \boldsymbol{w}_y, \kappa)$ is the p.d.f. of a vMF distribution,

$$\mathcal{V}(z^*; \boldsymbol{w}_y, \kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2}I_{n/2-1}(\kappa)}exp(\kappa \boldsymbol{w}_y^\top z^*), \quad (9)$$

where I_n denotes the first kind Bessel function at order n.

Within the vMF mixture model, the angular invariance induces a relationship between the target posterior $p^t(y|\mathbf{z})$ and the source mixture components $p(z^*|y)$.

Theorem 2. *When the assumptions 1 and 3 are satisfied, then*

$$p^t(y|z) = \frac{p(z^*|Y=y)w(r|\phi, d, y)P(Y=y)}{\sum_{c=1}^{C} p(z^*|Y=c)w(r|\phi, d, c)P(Y=c)}$$
$$= \frac{exp(\kappa \boldsymbol{w}_y^\top z^*)w(r|\phi, d, y)P(Y=y)}{\sum_{c=1}^{C} exp(\kappa \boldsymbol{w}_c^\top z^*)w(r|\phi, d, c)P(Y=c)}. \quad (10)$$

The eq. (10) in theorem 2 promises an optimization objective when there are enough observations for each source domain. However, the sample size is often prohibitive. Even if the sample complexity could be satisfied by the sample size, inputting too many source instances in mini-batches is not practical for a DNN. On the other hand, the empirical estimation $\hat{w}(r|\phi, d, c) = \hat{p}^t(\mathbf{z}|\phi, c)/\hat{p}(\mathbf{z}|\phi, d, c)$ goes to infinity when an instance $\mathbf{z}$ with polar coordinates (ϕ, r) is not observed in the c-th class and d-th domain ($\hat{p}(\mathbf{z}|\phi, d, c) = 0$).

The above practical concerns motivate us to modify eq. (10). For ease of illustration, we rewrite the vMF exponent factors $\mathbf{w}_c^\top \mathbf{z}^* \triangleq cos(\theta_c), c = 1, \ldots, C$. Considering the relative modification effect of the magnitude of observed empirical estimates $\hat{w}(r|\phi, d, y)$ on $cos(\theta_y)$, a finite $\hat{w}(r|\phi, d, y)$ endows a relatively (compared to an infinite empirical estimate $\hat{w}(r|\phi, d, c)$) small weight of the corresponding $cos(\theta_y)$, and thus a small credibility on θ_y. The proposed AIDGN reinterprets such credibility on θ_y as a perturbation to θ_y such that $exp(\kappa cos(\theta_y + \gamma w(r|\phi, d, y))) \approx exp(\kappa cos(\theta_y))w(r|\phi, d, y)$. For instances living on the mass manifold that are not observed, AIDGN reinterprets the infinite empirical estimate $\hat{w}(r|\phi, d, c)$ as absolute confidence on the corresponding θ_c, and no perturbation is added. In the light of reinterpreting $\hat{w}(r|\phi, d, \cdot)$ as perturbation on θ, enforcing $p^t(y|\mathbf{z})$ to be close to 1 reduces to enforcing

$$cos(\theta_y + \gamma w(r|\phi, d, y))p(y) \geq cos(\theta_c)p(c), \quad (11)$$

where $c \neq y$ is indexing any wrong class of $(\mathbf{z}, y)$. To derive the optimization objective of AIDGN, we adopt the maximum entropy principle again on the label distribution.

Theorem 3. *Let Δ be a $C - 1$ simplex and let $\boldsymbol{P} = (P_1, \ldots, P_C) \in \Delta$ denote a distribution of classes, then*

$$\boldsymbol{P}^* = (\frac{e^{\kappa(cos(\theta_1) - cos(\theta_y + \gamma w(r|\phi, d, y)))}}{\sum_{c=1}^{C} e^{\kappa(cos(\theta_c) - cos(\theta_y + \gamma w(r|\phi, d, y)))}},$$
$$\cdots, \frac{e^{\kappa(cos(\theta_C) - cos(\theta_y + \gamma w(r|\phi, d, y)))}}{\sum_{c=1}^{C} e^{\kappa(cos(\theta_c) - cos(\theta_y + \gamma w(r|\phi, d, y)))}})$$
$$= \arg\max_{\boldsymbol{P} \in \Delta} \kappa[\sum_{c \neq y}^{C} P_c(cos(\theta_c) - cos(\theta_y + \gamma w(r|\phi, d, y)))$$
$$+ P_y(cos(\theta_y + \gamma w(r|\phi, d, y) - cos(\theta_y + \gamma w(r|\phi, d, y)))]$$
$$+ \sum_{c=1}^{C} P_c(-logP_c).$$
$$(12)$$

At $\boldsymbol{P} = \boldsymbol{P}^*$, the maximum of the right hand side (r.h.s.) is

$$\ell_{AIDGN}(\boldsymbol{z}, y)$$
$$= -log\frac{e^{\kappa cos(\theta_y + \gamma w(r|\phi, d, y))}}{e^{\kappa cos(\theta_y + \gamma w(r|\phi, d, y))} + \sum_{c \neq y} e^{\kappa cos(\theta_y)}}. \quad (13)$$

In theorem 3 eq. (13), we derived the loss $\ell_{AIDGN}(\mathbf{z}, y)$. We next inspect the gradient behavior of the loss. We take the gradients of $\ell_{AIDGN}(\mathbf{z}, y)$ w.r.t. the correct center $\mathbf{w}_y$ and w.r.t. an incorrect center $\mathbf{w}_c$. At the k-th time step of training, let $\mathbf{p}_k$ be the prediction vector of the model,

$$\begin{cases} \partial\ell_{AIDGN}(\mathbf{z}, y)/\partial\mathbf{w}_y \propto \kappa(1 - p_k(y))\mathbf{z}, \\ \partial\ell_{AIDGN}(\mathbf{z}, y)/\partial\mathbf{w}_c \propto \kappa p_k(c)\mathbf{z}. \end{cases} \quad (14)$$

When performing the gradient descent training, the eq.(14) reveals the ideal behavior of decreasing gap between correct prediction probability and 1, and a non-ideal behavior in our DG formulation, i.e., shrinking the latent representation $\mathbf{z}$. Specifically, when we assume the angular invariance, the norm of $\mathbf{z}$ carries helpful information for comparing $p^t(\mathbf{z}|Y)$ and $p(\mathbf{z}|D, Y)$ as discussed in theorem 1, and the shrinking of $\mathbf{z}$ causes information loss in the norm coordinate.

To combat against the shrinking tendency brought by the $\ell_{AIDGN}(\mathbf{z}, y)$ loss, we introduce an information regularizer to penalize the loss of information. Concretely, we introduce an ideal norm distribution (which can be regarded as the norm distribution of an ideal source domain) that follows an exponential distribution $Exp(1/\mu^*)$. The information loss regularizer $KL(\cdot\|\cdot)$ is the Kullback-Leibler (K-L) divergence that calculates the relative entropy between $Exp(1/\mu^*)$ and $Exp(1/\mu)$.

$$KL(1/\mu^*\|1/\mu) = log(\mu/\mu^*) + \mu^*/\mu - 1 \quad (15)$$

Finally, the final optimization objective of the proposed AIDGN is

$$L_{AIDGN} = \sum_{d=1}^{N} \sum_{i=1}^{n_d} \eta KL(1/\mu^*\|1/\mu_{d,y,\phi})$$
$$- log\frac{e^{\kappa cos(\theta_y + \gamma w(r|\phi, d, y))}}{e^{\kappa cos(\theta_y + \gamma w(r|\phi, d, y))} + \sum_{c \neq y} e^{\kappa cos(\theta_y)}}, \quad (16)$$

where $\eta > 0$ is a hyperparameter controlling the trade-off between information memorizing and forgetting.

Method	PACS	VLCS	OfficeHome	TerraIncognita	Avg
ERM [Vapnik, 1999]	85.5 ± 0.2	77.5 ± 0.4	66.5 ± 0.3	46.1 ± 1.8	68.9
ERM† [Vapnik, 1999]	85.8 ± 0.2	77.7 ± 0.4	66.9 ± 0.2	45.8 ± 1.5	69.1
IRM [Arjovsky *et al.*, 2020]	83.5 ± 0.8	78.5 ± 0.5	64.3 ± 2.2	47.6 ± 0.8	68.5
DRO [Sagawa *et al.*, 2019]	84.4 ± 0.8	76.7 ± 0.6	66.0 ± 0.7	43.2 ± 1.1	67.6
Mixup [Wang *et al.*, 2020b]	84.6 ± 0.6	77.4 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	69.5
MLDG [Li *et al.*, 2018a]	84.9 ± 1.0	77.2 ± 0.4	66.8 ± 0.6	47.7 ± 0.9	69.2
CORAL [Sun and Saenko, 2016]	86.2 ± 0.3	78.8 ± 0.6	68.7 ± 0.3	47.6 ± 1.0	70.3
MMD [Li *et al.*, 2018b]	84.6 ± 0.5	77.5 ± 0.9	66.3 ± 0.1	42.2 ± 1.6	67.7
DANN [Ganin *et al.*, 2016]	83.6 ± 0.4	78.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	68.7
CDANN [Li *et al.*, 2018c]	82.6 ± 0.9	77.5 ± 0.1	65.8 ± 1.3	45.8 ± 1.6	67.9
MTL [Blanchard *et al.*, 2021]	84.6 ± 0.5	77.2 ± 0.4	66.4 ± 0.5	45.6 ± 1.2	68.5
SagNet [Nam *et al.*, 2021]	86.3 ± 0.2	77.8 ± 0.5	68.1 ± 0.1	48.6 ± 1.0	70.2
ARM [Zhang *et al.*, 2020]	85.1 ± 0.4	77.6 ± 0.3	64.8 ± 0.3	45.5 ± 0.3	68.3
VREx [Krueger *et al.*, 2021]	84.9 ± 0.6	78.3 ± 0.2	66.4 ± 0.6	46.4 ± 0.6	69.0
RSC [Huang *et al.*, 2020]	85.2 ± 0.9	77.1 ± 0.5	65.5 ± 0.9	46.6 ± 1.0	68.6
AIDGN (ours)	$\mathbf{86.6 \pm 0.3}$	$\mathbf{78.9 \pm 0.3}$	$\mathbf{68.8 \pm 0.2}$	$\mathbf{49.4 \pm 0.6}$	$\mathbf{70.9}$

Table 1: **Benchmark Comparisons.** Out-of-domain classification accuracies(%) on PACS, VLCS, OfficeHome and TerraIncognita are shown. Note that the results of ERM† are reproduced by us, and other numbers are from DomainBed.

Remark 1. *The proposed AIDGN optimization objective eq. (16) theoretically justifies and generalizes the recent proposed MAG loss for face recognition [Meng* et al.*, 2021]. When there is only one source domain, it can be easily verified that with a first-order Maclaurin's expansion approximation w.r.t. the log term in the regularizer $KL(1/\mu^* \| 1/\mu)$, the proposed AIDGN loss eq. (16) degenerates to the MAG loss.*

4 Experiments

4.1 Experimental Settings

Datasets. We conduct our experiments[3] on four public benchmark datasets to evaluate the effectiveness of the proposed AIDGN. PACS [Li *et al.*, 2017] comprises four domains $d \in \{$photo, art, cartoon, sketch$\}$, containing 9991 images of 7 categories. VLCS [Fang *et al.*, 2013] comprises four photographic domains $d \in \{$VOC2007, LabelMe, Caltech101, SUN09$\}$, with 10729 samples of 5 classes. Office-Home [Venkateswara *et al.*, 2017] has four domains $d \in \{$art, clipart, product, real$\}$, containing 15500 images with a larger label sets of 65 categories. TerraIncognita [Beery *et al.*, 2018] comprises photos of wild animals taken by cameras at different locations. Following [Gulrajani and Lopez-Paz, 2021], we use domains of $d \in \{$L100, L38, L43, L46$\}$, which include 24778 samples and 10 classes.

Evaluation Protocol. For a fair comparison, we use the DomainBed training-domain validation set protocol [Gulrajani and Lopez-Paz, 2021] for model selection. For training, we randomly split each training domain into 8:2 training/validation splits, choose the model on the overall validation set, and then evaluate its performance on the target domain set. We report the mean and standard deviation of out-of-domain classification accuracy from three different runs with different training-validation splits.

Implementation Details. For all datasets, we use ResNet-50 [He *et al.*, 2016] pre-trained on ImageNet [Deng *et al.*,

---

[3]Codes are avalable at https://github.com/JinYujie99/aidgn

2009] as the feature extractor π and one fully connected layer as the classifier f. We construct a mini-batch containing all source domains where each domain has 32 images. We freeze all the batch normalization (BN) layers from pre-trained ResNet since different domains in a mini-batch follow different distributions. The network is trained for 5000 iterations using the Adam [Kingma and Ba, 2015] optimizer. To estimate $\delta_{y,\phi}$ and $\mu_{d,y,\phi}$ in AIDGN loss (16), we treat all $\delta_{y,\phi}$ as one single hyperparameter δ, and perform in-batch estimation for the d-th domain norm scale parameter, i.e., μ_d, while ignoring the index y and ϕ. Specifically, we estimate μ_d by computing the average norm of samples of the d-th domain in a minibatch. We do this mainly for two reasons. First, according to the norm shift assumption, domain index d is more relevant to the norm distribution. Secondly, since the feasible label sets and angular sets can be extensive, it is almost impossible to estimate $\mu_{d,y,\phi}$ in a minibatch precisely. Moreover, for the log term in the KL regularizer, we approximate it with a first-order Maclaurin's expansion to make the overall objective function convex. Following [Gulrajani and Lopez-Paz, 2021], we conduct a random search over the joint distribution of hyperparameters. More implementation details about data preprocessing techniques, model architectures, hyperparameters, and experimental environments can be found in the supplementary material.

4.2 Benchmark Comparisons

We compare our proposed AIDGN with 14 available DG methods in DomainBed [Gulrajani and Lopez-Paz, 2021]. Results on PACS, VLCS, OfficeHome, and TerraIncognita are reported in Table 1. (*Details of each baseline and full results per dataset per domain can be found in the supplementary material.*) All the values of baselines are taken from DomainBed when ResNet-50 is used as a backbone network, except that ERM† is reproduced by us. For all datasets, AIDGN achieves better average out-of-domain accuracy. Particularly, in TerraIncognita, the proposed method achieves 49.4%, which is significantly better than the most competitive baseline, 48.6%.

	C	L	V	S	Avg
AIDGN w/o RD	96.9 ± 0.3	63.5 ± 0.5	73.0 ± 0.5	75.8 ± 0.6	77.3
AIDGN w/o R	97.3 ± 0.5	64.9 ± 0.8	$\mathbf{73.4 \pm 0.4}$	77.9 ± 1.0	78.4
AIDGN w/o D	97.6 ± 0.3	64.7 ± 0.4	$\mathbf{73.4 \pm 0.4}$	78.1 ± 0.7	78.4
AIDGN	$\mathbf{98.3 \pm 0.2}$	$\mathbf{65.7 \pm 0.4}$	73.1 ± 0.4	$\mathbf{78.7 \pm 0.7}$	$\mathbf{78.9}$

Table 2: **Ablation studies.** Out-of-domain classification accuracies(%) with different optimization components on VLCS.
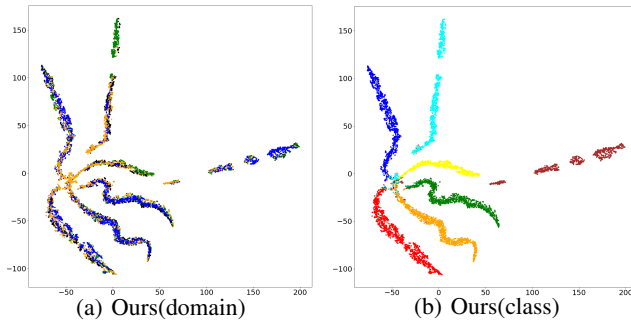

(a) Ours(domain)   (b) Ours(class)

Figure 2: Feature visualization for AIDGN on PACS: (a) different colors indicate different domains, source domains include cartoon (black), photo (green) and sketch (blue) while the target domain is art-painting (orange); (b) different colors represent different classes. *Best viewed in color(Zoom in for details).*

Similarly, it is 86.6% for PACS, 78.9% for VLCS, and 68.8% for OfficeHome, all of which outperform the previous best domain-invariant representation learning results.

4.3 Ablation Studies

To verify the effectiveness of all components of the AIDGN optimization objective, we do ablation studies on the VLCS dataset. Specifically, we compare AIDGN with three variants: a) AIDGN w/o R: The model is trained without the KL regularizer. b) AIDGN w/o D: The model is trained without estimation for the norm scale parameters of source domains, i.e., ignoring the μ_d term in the expression of $w(r|\phi, d, y)$. c) AIDGN w/o RD: The model is trained without both. The results are reported in Table 2. Note that the results for all the three variants also follow the training-domain validation set model selection and hyperparameter random search protocol, and are obtained from three different runs with different dataset splits. The results show that the full model AIDGN outperforms all these variants, indicating that both the KL regularizer and the estimation for the source domain norm scale parameters are essential to our algorithm.

4.4 Discussions on the AIDGN Latent Space

To analyze the learned latent space of AIDGN, we visualize the distributions of features with t-SNE in Fig 2. It is shown that AIDGN can effectively make the decision boundaries of all categories separate from each other and better balance the intra-class compactness and the inter-class separation.

In addition, we test our model with existing test-time adaptation methods which are based on entropy minimization: T3A [Iwasawa and Matsuo, 2021] and Tent [Wang *et al.*,
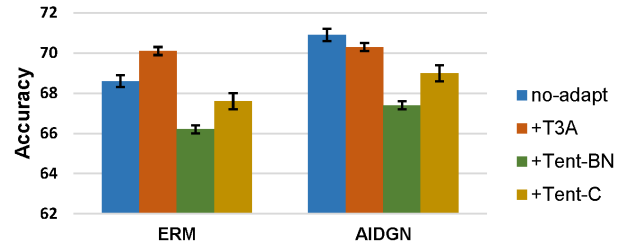

Figure 3: Average out-of-domain accuracies of ERM and AIDGN equipped with different test-time adaptation methods.

2020a]. Since DomainBed [Gulrajani and Lopez-Paz, 2021] freezes the BN layers, we test two slightly modified versions of Tent following [Iwasawa and Matsuo, 2021]. Specifically, Tent-BN adds one BN layer before the classifier and modulates its transformation and normalization parameters. Tent-C adapts the classifier to minimize prediction entropy. The average performances across PACS, VLCS, OfficeHome and TerraIncognita are shown in Fig 3. (Please refer to the full per dataset and domain results in the supplementary material.) We can find that none of them can further improve the performance of AIDGN and our original AIDGN outperforms all the variants. While test-time adaptation modifies the model aiming at reducing prediction uncertainty caused mainly by domain shift, our proposed AIDGN models domain shift as norm parameters shift in the latent space, and is thus robust to the uncertainty brought by domain shift.

5 Conclusion

In this paper, we introduce a novel angular invariance and norm shift assumption into domain generalization, inspired by the acknowledged fact that internal layers in convolutional neural networks capture high-level semantic concepts. We then propose a method based on the assumption and develop a practical optimization objective within a von-Mises Fisher mixture model. Extensive experiments on four benchmarks demonstrate the superior performance of our proposed method. While this work focuses on inter-domain invariance, it is complementary to ensemble learning which aims to learn robust classifiers. For future work, we will develop methods that take both invariance and robustness into consideration.

Acknowledgements

References

[Arjovsky *et al.*, 2020] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *stat*, 1050:27, 2020.

[Beery *et al.*, 2018] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, pages 456–473, 2018.

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. In *Neurips*, volume 19, 2007.

[Blanchard *et al.*, 2011] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Neurips*, volume 24, 2011.

[Blanchard *et al.*, 2021] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *JMLR*, 22(2), 2021.

[Blumenson, 1960] LE Blumenson. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67(1), 1960.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Fang *et al.*, 2013] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pages 1657–1664, 2013.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1), 2016.

[Gopal and Yang, 2014] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *ICML*, pages 154–162, 2014.

[Gulrajani and Lopez-Paz, 2021] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2020] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140, 2020.

[Iwasawa and Matsuo, 2021] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Neurips*, volume 34, 2021.

[Jaynes, 1957] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4), 1957.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neurips*, volume 25, 2012.

[Krueger *et al.*, 2021] 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. In *ICML*, pages 5815–5826, 2021.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[Li *et al.*, 2018a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 5542–5550, 2018.

[Li *et al.*, 2018b] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2018c] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, volume 32, 2018.

[Liu *et al.*, 2021] Xiaofeng Liu, Bo Hu, Linghao Jin, Xu Han, Fangxu Xing, Jinsong Ouyang, Jun Lu, Georges EL Fakhri, and Jonghye Woo. Domain generalization under conditional and label shifts via variational bayesian inference. In *IJCAI*, 2021.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Meng *et al.*, 2021] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *CVPR*, pages 14225–14234, 2021.

[Nam *et al.*, 2021] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, pages 8690–8699, 2021.

[Sagawa *et al.*, 2019] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint:1911.08731*, 2019.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 1999.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[Wang *et al.*, 2020a] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2020.

[Wang *et al.*, 2020b] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP*, pages 3622–3626, 2020.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[Zhang *et al.*, 2020] 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:2007.02931*, 2020.

[Zhao *et al.*, 2019] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *ICML*, pages 7523–7532, 2019.