Generalizability of Adversarial Robustness Under Distribution Shifts

Anonymous submission

Abstract

Recent progress in empirical and certified robustness promises to deliver reliable and deployable Deep Neural Networks (DNNs). Despite that success, most existing evaluations of DNN robustness have been done on images sampled from the same distribution that the model was trained on. Yet, in the real world, DNNs may be deployed in dynamic environments that exhibit significant distribution shifts. In this work, we take a first step towards thoroughly investigating the interplay between empirical and certified adversarial robustness on one hand and domain generalization on another. To do so, we train robust models on multiple domains and evaluate their accuracy and robustness on an unseen domain. We observe that: (1) both empirical and certified robustness generalize to unseen domains, and (2) the level of generalizability does not correlate well with input visual similarity, measured by the FID between source and target domains. We also extend our study to cover a real-world medical application, in which adversarial augmentation enhances both the robustness and generalization accuracy in unseen domains.

1 Introduction

Deep Neural Networks (DNNs) are vulnerable to small and carefully designed perturbations, known as adversarial attacks (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). That is, a DNN $f_\theta : \mathbb{R}^d \to \mathcal{P}(\mathcal{Y})$ can produce two different predictions for the inputs $x \in \mathbb{R}^d$ and $x + \delta$, although both x and $x + \delta$ are perceptually indistinguishable. Furthermore, DNNs are found to be brittle against simple semantic transformations such as rotation, translation, and scaling (Engstrom et al. 2019). These observations raised concerns regarding the deployability of DNNs in security-critical applications, such as self-driving and medical diagnosis (Papernot et al. 2016; Finlayson et al. 2019; Ma et al. 2021). This brittleness provoked several efforts to build models that are not only accurate but also *robust* (Gu and Rigazio 2015). Building robust models is usually achieved either **(i)** *empirically*, where the DNN's training routine is changed to include such malicious adversarial examples in the training set (Madry et al. 2018), or **(ii)** *certifiably*, where theoretical guarantees are given about the robustness of a DNN in a region around a given input x (Lécuyer et al. 2019).

Despite great progress in the adversarial robustness literature on building accurate and robust models, most approaches are tested on *in-distribution* data. In other words, the scenario considered is one in which both the training and testing sets are independently and identically distributed (IID). However, this IID assumption rarely holds in practice, as data in the real world can be sampled from various distributions with significant domain shifts. For example, a deep-learning based medical image classifier may be trained on data collected from one hospital, but later deployed in a different hospital (Bándi et al. 2019). Unfortunately, DNNs struggle to generalize to out-of-domain data (Geirhos et al. 2020, 2021), even in the absence of of adversarial examples. This lack of generalization has led the research community to invest in the problem of Domain Generalization (DG). The aim of DG is to learn invariant representations from diverse distributions of data, denoted as *source* domains, such that these representations generalize to an unseen distribution, known as the *target* domain (Wang et al. 2021; Gulrajani and Lopez-Paz 2021). This setup mimics the unexpected nature of real-world distribution shifts, where models are constantly exposed to novel domains, and fine-tuning on all these domains becomes impractical. While there has been effort in improving the generalization of DNNs (Tzeng et al. 2014; Sun and Saenko 2016; Motiian et al. 2017; Zhang et al. 2021; Shen et al. 2021; Wang et al. 2021; Zhou et al. 2022), the generalizability of adversarial robustness to unseen domains remains unexplored.

In this work, we set out to study the interplay between domain generalization and adversarial robustness. We conduct comprehensive experimental studies on five standard DG benchmarks provided by DomainBed (Gulrajani and Lopez-Paz 2021) and WILDS (Koh et al. 2021). In our experiments, we study both empirical and certified robustness against input perturbations and spatial deformations. We first investigate the generalizability of empirical robustness, which a DNN obtains by employing the popular adversarial training method (Madry et al. 2018) while training on the source data. We find that, in many scenarios, improving empirical robustness in the source domain generalizes to the target domain with little cost on the performance of the model on unperturbed data. We then inspect the generalizability of certified robustness against both input perturbations and parametric deformations by employing Randomized Smoothing (RS) (Cohen, Rosenfeld, and Kolter 2019) and DeformRS (Alfarra et al. 2022a). We observe that cer-

tified robustness generalizes to unseen domains when using randomized smoothing frameworks against pixel perturbations and five different input deformations including rotation, translation, and scaling. To the best of our knowledge, we provide the first large scale experimental analysis of the generalizability of adversarial robustness to unseen domains. Our analysis leads to the following contributions:

1. We contrast the behavior of robustness under both transfer learning and domain generalization. Unlike transfer learning, domain generalization does not always improve through robust training.
2. We empirically show that visual similarity, between the source and target domains, does not correlate well with the level of generalizability to the target domain.
3. We analyze a practical medical application, in which adversarial training in the source domain improves the generalization of accuracy and robustness in the unseen target domain.

2 Related Work

Domain Generalization. Domain generalization (DG) studies the ability of models to learn representations that can be readily applied to data from unseen domains (Wang et al. 2021; Zhou et al. 2022). In the DG setup, a model is trained on multiple source domains and then evaluated on an unseen target domain, which exhibits a significant shift from the training domains. DG approaches can be categorized into different groups. (i) Data augmentation techniques (Gong et al. 2019; Zhou et al. 2020, 2021), (ii) Representation learning methods (Blanchard, Lee, and Scott 2011; Nguyen et al. 2021; Lu et al. 2022), and (iii) Learning-strategy approaches (Li et al. 2018; Carlucci et al. 2019; Cha et al. 2021). In this work, we study DG from an adversarial robustness lens. In particular, we analyze both the generalization accuracy and robustness of adversarially trained classifiers on unseen domains.

Adversarial Robustness. Adversarial attacks are imperceptible, semantic-preserving perturbations that can fool DNNs (Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2014). Given the security concerns that adversarial attacks induced, several works proposed changing the training routine to enhance model robustness (Madry et al. 2018; Zhang et al. 2019). While empirical defenses are effective in enhancing the robustness of the underlying model, such approaches do not guarantee robustness. As a result, there has been a growing interest in certifiably robust classifiers, for which no adversary can exist in a specified region around a data point (Raghunathan, Steinhardt, and Liang 2018; Mohapatra et al. 2020; Lee et al. 2021). A scalable approach to achieving certified robustness is Randomized Smoothing (RS) (Cohen, Rosenfeld, and Kolter 2019). RS constructs a smooth classifier from any arbitrary base classifier by outputting the most probable class when the input is subjected to Gaussian noise. Recently, DeformRS extended RS to provide certified robustness against parameterized geometric deformations (Alfarra et al. 2022a; S. et al. 2022). In this work, we set out to study the interplay between (empirical and certified) robustness and domain generalization by deploying adversarial training, RS, and DeformRS.

Adversarial Training in Dynamic Environments. To improve the ability of machine learning models to learn generalizable knowledge, researchers have proposed several problems, such as transfer learning, continual learning, domain adaptation, and domain generalization (Zhuang et al. 2020; Delange et al. 2021; Wang and Deng 2018; Wang et al. 2021). Among these problems, only transfer learning, where a model pre-trains on tasks with large datasets and then adapts to downstream tasks with limited data, has been thoroughly studied under the lens of adversarial robustness. (Salman et al. 2020) showed that, in terms of downstream task accuracy, adversarially trained representations outperform nominally trained representations. (Utrera et al. 2021) further explained that adversarial training in the source domain increases shape bias, resulting in better transferability. Finally, (Deng et al. 2021) provided theoretical justification to support these empirical findings. Besides downstream task accuracy, (Shafahi et al. 2020) studied the transferability of robustness itself. Although useful, these transfer learning results presume fine-tuning on the target domain, which is infeasible in many real-life scenarios. Table 1 illustrates the differences between transfer learning and domain generalization, which is the setup we adopt. In this paper, we take a first step to empirically investigate whether adversarial training leads to robust representations that generalize without requiring prior knowledge of the target domain.

3 Background on Domain Generalization

Domain Generalization Setup. Given an input space $\mathcal{X}$ and a label space $\mathcal{Y}$, one can define a joint distribution $\mathbb{P}_{XY}$ over $\mathcal{X}$ and $\mathcal{Y}$. A domain, or distribution, is a collection of samples drawn from $\mathbb{P}_{XY}$. In our setting, the input and label spaces are fixed, but we may have multiple unique joint distributions. Specifically, we assume that there are N source domains of varying sizes. For each task n, $S_n = \{(x_j, y_j)\}_{j=1}^{|S_n|} \sim \mathbb{P}_{XY}^{(n)}$. We denote the training set by $S = \{S_i | i = 1, \ldots, N\}$. The aim of DG is to use S to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the error on some unseen target domain $T \sim \mathbb{P}_{XY}^{(N+1)}$. We enforce that $\mathbb{P}_{XY}^{(k)} \neq \mathbb{P}_{XY}^{(n)}$ for $k \neq n, k, n \in \{1, \ldots, N+1\}$, which means that the target domain is distinct from the source domains that are, in turn, also distinct from each other. More formally, given S, we seek a parameterized model f_{θ^*} such that:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{P}_{XY}^{(T)}} \left[\mathcal{L}(f_\theta(x), y)\right], \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss for the classification task. Note that the model is not allowed to sample the target domain during training, so most methods use the empirical risk of the source datasets as a proxy for the true target risk. The supervised average risk ($\mathcal{E}$) is given by:

$$\mathcal{E} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{|S_n|} \sum_{i=1}^{|S_n|} [\mathcal{L}(f_\theta(x), y)] \tag{2}$$

Problem Setup	Training Data	Target Data	Problem Condition	Access to Target
Transfer learning	S_{source}, S_{target}	S_{target}	$\mathcal{Y}_{source} \neq \mathcal{Y}_{target}$	✓
Domain generalization	$S = \{S_i \vert i = 1, \ldots, N\}$	S_{N+1}	$\mathbb{P}_{XY}(S_k) \neq \mathbb{P}_{XY}(S_n)$ for $k \neq n$	✗

Table 1: **Comparison between Domain Generalization (DG) and Transfer Learning.** DG differs from transfer learning in two ways. 1) The model in DG never sees the target data during training, so fine-tuning on the target is not allowed. 2) The target labels are kept fixed in DG; however, the target samples are drawn from a domain that is distinct from the source domains.

with $(x, y) \sim S$. In practice, we define a fixed held-out validation set $S^v \subset S$. The average risk on this source validation set is used to select the best model, which is evaluated on the target domain without any fine-tuning steps. In what follows, Section 4 (and Section 5) investigates the generalizability of empirical (and certified) robustness to diverse target domains.

4 Empirical Robustness and Domain Generalization

In this section, we study the generalizability of empirical robustness methods that enhance the adversarial robustness of DNNs. We begin with a brief introduction of Adversarial Training (AT) (Madry et al. 2018), after which we study the effect of deploying AT in a domain generalization setup.

4.1 Background and Setup

Adversarial Attacks. Adversarial attacks are small imperceptible perturbations that, once added to a "clean" input sample, cause the classifier f_θ to misclassify the perturbed sample. Formally, let (x, y) be an input label pair where f_θ correctly classifies x (*i.e.* $\arg\max_i f^i_\theta(x) = y$). An attacker crafts a small perturbation δ such that $\arg\max_i f^i_\theta(x+\delta) \neq y$, which is usually obtained by solving the following optimization problem:

$$\max_{\delta} \ \mathcal{L}(f_\theta(x+\delta), y) \qquad \text{s.t. } \|\delta\|_p \leq \epsilon, \tag{3}$$

where $p \in \{2, \infty\}$, $\epsilon > 0$ is a small constant that enforces the imperceptibility of the added perturbation, and $\mathcal{L}$ is a suitable loss function (*e.g.* Cross Entropy). Let δ^* be the solution to the problem in Eq. 3, then the adversarial example is denoted by $x_{adv} = x + \delta^*$.

Adversarial Training as Augmentation. Adversarial Training (AT) (Madry et al. 2018) trains the classifier on adversarial examples rather than clean samples. In particular, AT obtains the network parameters θ^* by solving the following optimization problem:

$$\min_{\theta} \ \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta, \|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(x+\delta), y) \right], \tag{4}$$

where $\mathcal{D}$ is a data distribution. In general, the inner maximization problem is solved through K steps of Projected Gradient Descent (PGD) (Madry et al. 2018). While conducting adversarial training enhances the model's robustness against adversarial attacks, this usually comes at the cost of losing some clean accuracy (performance on unperturbed samples). To alleviate the drop in performance, we follow the method by (Zhang et al. 2019) and deploy adversarial training as a data augmentation scheme. In particular, we obtain network parameters θ^* that minimize the following objective:

$$\theta^* = \arg\min_{\theta} \ \mathbb{E}_{(x,y)\sim\mathbb{P}^{(T)}_{XY}} \left[\lambda \mathcal{L}(f_\theta(x), y) + (1-\lambda)\mathcal{L}(f_\theta(x_{adv}), y) \right], \tag{5}$$

where $\lambda \in [0, 1]$ controls the robustness-accuracy trade-off.

Training & Evaluation Setup. In our experiments, we focus on image classification and adopt the framework of DomainBed (Gulrajani and Lopez-Paz 2021), which is the standard benchmark in the image domain generalization literature. All models are initialized with a ResNet-50 backbone pre-trained on ImageNet-1K (Deng et al. 2009). We train all models with adversarial augmentation to minimize the objective in Eq. 5 on the source domains, where x_{adv} is computed with a Projected Gradient Descent (PGD) attack (Madry et al. 2018) using 20 PGD steps. The target domain remains unseen until test time. Specifically, we follow the training-domain validation strategy described in DomainBed for model selection. We experiment with a range of perturbation budgets (ϵ) on various datasets: PACS, Office-Home, VLCS, and TerraIncognita (Gulrajani and Lopez-Paz 2021). We report the ℓ_∞ results in the main paper, where we use $\epsilon \in \{0, {}^8\!/_{255}, {}^{16}\!/_{255}, {}^{32}\!/_{255}\}$. The appendix includes experiments for ℓ_2 perturbations due to space constraints. Note that for $\epsilon = 0$, the training objective reduces to the empirical risk minimization in Eq. (2). Each model is trained on one value of ϵ but is evaluated on all four values of ϵ with 20 steps of PGD attacks. In the following experiments, we fix λ in Eq. 5 to 0.5 and leave the ablation to the appendix.

4.2 The Generalization of Empirical Robustness

In this section, we investigate the generalizability of empirical robustness to unseen domains. More precisely, we are interested in understanding the interplay between standard accuracy and robust accuracy in the scope of source vs. target domains. We report in Figure 1 the standard accuracy (first column of each matrix) and robust accuracy against different values of ϵ on all considered datasets. We further summarize the clean and robust accuracy at $\epsilon = {}^8\!/_{255}$ in Table 2 for ease of comparison. Next, we analyze these results to answer the following questions.

Q1: Do adversarially robust models generalize better than their standard-trained counterparts? No, which is evident from the first two rows of Table 2. With the exception of the smallest dataset PACS, the clean accuracy of the robust model in the unseen domain is lower than that of the standard-trained model by 3.6% or more. ❶ **Unlike transfer learning, where robust training in the source domain**

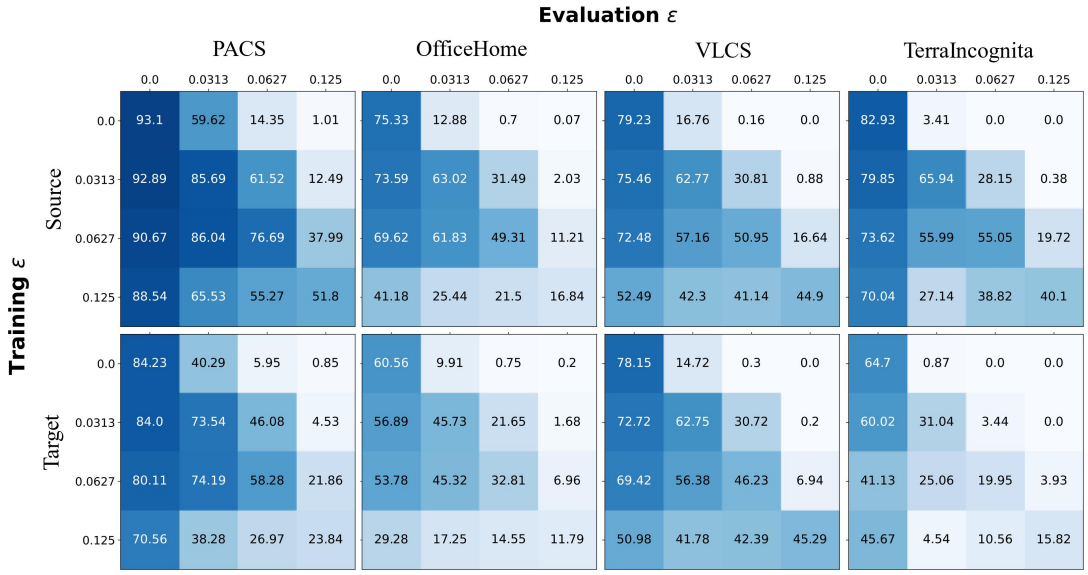


Figure 1: **Evaluation of ℓ_∞ Robustness.** We train models in the source domain and evaluate them in both the source and target domains with different ϵ values. In the Source tables, the $(i^{\text{th}}, j^{\text{th}})$ entry represents a model trained with $\epsilon = i$ and evaluated on $\epsilon = j$ on samples drawn from source domain. In the Target tables, the $(i^{\text{th}}, j^{\text{th}})$ entry is a model trained in the source with $\epsilon = i$ and evaluated on $\epsilon = j$ on samples drawn from target domain. No fine-tuning is done in the target domain.

		Dataset			
		PACS	**OfficeHome**	**VLCS**	**TerraIncognita**
Clean Accuracy (%)	**Standard Model**	84.23	60.56	78.15	64.70
	Robust Model	84.00	56.89	72.72	60.02
Robust Accuracy (%)	**Standard Model**	40.29	9.91	14.72	0.87
	Robust Model	73.54	45.73	62.75	31.04

Table 2: **Generalization of Robustly- and Nominally- Trained Models on Various Datasets.** Applying adversarial training on the source domain leads to significant improvements in the model's robustness in the target domain, relative to the observed drop in the standard accuracy.

is favorable, robust training does not improve generalization to the target domain if no fine-tuning is allowed. This result contrasts with findings from the transfer learning literature, where models trained robustly in the source domain outperform standard-trained models across a variety of downstream tasks (Salman et al. 2020). It is especially surprising given that previous works suggest that robust training encourages shape bias over texture bias, hinting at better generalization (Geirhos et al. 2019; Utrera et al. 2021). Moreover, (Deng et al. 2021) showed that adversarial training in the source domain results in provably better representations for fine-tuning on the target domain. Such seemingly contradictory findings can be reconciled by considering the key differences between transfer learning and domain generalization summarized in Table 1. Specifically, previous works in transfer learning assume that the model can sample the target domain at some point to perform fine-tuning. Since domain generalization does not allow access to the target domain, such benefits are not guaranteed. *We encourage future works to investigate under what conditions adversarial training helps the generalization accuracy with no fine-tuning on the target.*

Q2: Does a higher source domain robustness correspond to a higher target domain robustness? As expected, DNNs lose some robustness when evaluated on a target domain that is distinct from the training domains. This observation is evident by comparing any cell in the top row (Source) tables in Figure 1 with the corresponding cell in the second row (Target) tables. For example, the TerraIncognita model, which is trained and evaluated on ($\epsilon = 8/255$) adversaries, loses around 35% accuracy when the distribution shifts to the target domain. Nevertheless, by observing that all the source and target tables have similar color trends, we find that ❷ **higher robustness in the source domain corresponds to higher robustness in the target domain.** Our results suggest that one way to increase the out-of-distribution robustness of a deployed model is to improve its robustness in the source validation set, which supports the applicability of ongoing efforts in adversarial robustness research (Zhang et al. 2019; Wang et al. 2020; Wu, Xia, and Wang 2020).

Q3: Does the robustness-accuracy trade-off generalize to unseen domains? As observed in Figure 1, ❸ **achieving a more robust model comes at the cost of standard accuracy not only in the source domain, but also in the target**

domain. Looking at OfficeHome, the target robust accuracy of a robust model (trained and evaluated on $\epsilon = 16/255$) is 50% more than that of the standard-trained model. Yet, the clean accuracy of the robust model is about 6% less than the standard model accuracy. In general, as the network becomes more robust to a particular perturbation budget ϵ in the source domain, it becomes more robust to adversaries generated within that budget in the target domain. Nevertheless, the performance of the robust network on clean samples decreases in both domains. Therefore, *consistent with the robustness literature (Tsipras et al. 2019), robustness comes at the cost of standard accuracy even in the unseen target domains.*

5 Certified Robustness and Domain Generalization

In Section 4, we analyzed the interplay between empirical robustness (obtained by adversarial training) and domain generalization. While empirical robustness studies give hints about the reliability of a given model under adversarial attacks, they give no guarantees against the existence of such adversaries. To deploy DNNs in dynamic environments (Koh et al. 2021), we need robustness guarantees to carry over into unseen domains. To that end, we study the generalizability of the certified robustness of DNNs. We deploy Randomized Smoothing (RS) and DeformRS to certify DNNs against input perturbations and deformations. We start by giving a brief overview of RS and DeformRS.

5.1 Background and Setup

Certifying Against Additive Perturbations and Input Deformations. Randomized smoothing (RS) (Cohen, Rosenfeld, and Kolter 2019) is a method for constructing a "smooth" classifier from a given classifier f_θ. The smooth classifier returns the average prediction of f_θ when the input x is subjected to additive Gaussian noise:

$$g_\theta(x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}\left[f_\theta(x+\epsilon)\right]. \tag{6}$$

Let g_θ predict label c_A for input x with some confidence, *i.e.* $\mathbb{E}_\epsilon[f_\theta^{c_A}(x+\epsilon)] = p_A \geq p_B = \max_{c\neq c_A}\mathbb{E}_\epsilon[f_\theta^{c}(x+\epsilon)]$, then, as shown by (Zhai et al. 2020), g_θ's prediction is certifiably robust at x with certification radius:

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right), \tag{7}$$

where Φ^{-1} is the inverse CDF of the standard Gaussian distribution. As a result of Eq. 7, $\arg\max_i g_\theta^i(x+\delta) = \arg\max_i g_\theta^i(x),\ \forall\|\delta\|_2 \leq R$.

While Eq. 7 provides theoretical guarantees for robustness against additive perturbations, DNNs are also brittle against simple input transformations such as rotation. (Alfarra et al. 2022a) extended randomized smoothing to certify parametric input deformations through DeformRS, which defined the parametric smooth classifier for a given input x with pixel coordinates p as follows:

$$g_\phi(x,p) = \mathbb{E}_{\epsilon\sim\mathcal{D}}\left[f_\theta(I_T(x, p+\nu_{\phi+\epsilon}))\right], \tag{8}$$

where I_T is an interpolation function (*e.g.* bilinear interpolation) and ν_ϕ is a parametric deformation function with parameters ϕ (*e.g.* ν is a rotation function and ϕ is the rotation angle). Analogous to the RS formulation in Eq. 6, g outputs the average prediction of f_θ over deformed versions of the input x. (Alfarra et al. 2022a) showed that parametric-domain smooth classifiers are certifiably robust against perturbations to the parameters of the deformation function. In particular, g's prediction is constant with certification radius:

$$\begin{aligned} R &= \sigma\left(p_A - p_B\right) && \text{for } \mathcal{D} = \mathcal{U}[-\sigma,\sigma], \\ R &= \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right) && \text{for } \mathcal{D} = \mathcal{N}(0,\sigma^2 I). \end{aligned} \tag{9}$$

Put simply, as long as the perturbations to the deformation function parameters (*e.g.* rotation angle) are within R, the prediction of g remains constant. In this work, we leverage RS and DeformRS to study the generalizability of certified robustness to unseen target domains.

Experimental Setup. To split the data into source and target domains, we use the *Photo, Art, Cartoon,* and *Sketch* distributions from PACS (Li et al. 2017). We use RS to certify pixel perturbations and DeformRS to certify five input deformations: rotation, translation, scaling, affine, and a deformation characterized by a Discrete Cosine Transform (DCT) basis. Following (Gulrajani and Lopez-Paz 2021), we employ data augmentation during training and train solely on the source domains. To evaluate the certified robustness of the trained classifier, we plot the certified accuracy curves for both the source and target domains for each considered deformation. The certified accuracy at a radius R is the percentage of the test set that is both classified correctly and has a certified radius of at least R. We calculate the certified radius for a given input through either Eq. 7 for pixel perturbations or Eq. 9 for input deformations. Here, we report the envelope plots, which illustrate the best certified accuracy per radius over all values of the smoothing deformation parameter ϕ. We leave the detailed results for each choice of ϕ to the appendix. We employ Monte Carlo sampling with 100k samples to estimate p_A and bound $p_B = 1 - p_A$ by following the standard practice (Zhai et al. 2020; Cohen, Rosenfeld, and Kolter 2019; Alfarra et al. 2022a). Finally, we follow (Zhai et al. 2020) in reporting the Average Certified Radius (ACR) of correctly classified samples.

Regarding the architecture, we follow the DomainBed (Gulrajani and Lopez-Paz 2021) benchmark in selecting ResNet-50 as a backbone. To assess the effect of deploying a more powerful architecture on the generalizability to unseen domains, we further include experiments with the recent transformer model ViT-Base (Dosovitskiy et al. 2021).

5.2 Generalizability of Certified Robustness to Unseen Target Domains

We investigate under what scenarios the certified robustness generalizes to unseen domains. We first show how much certified accuracy (CA) is maintained when the target domain exhibits a distribution shift. Then, we study whether a

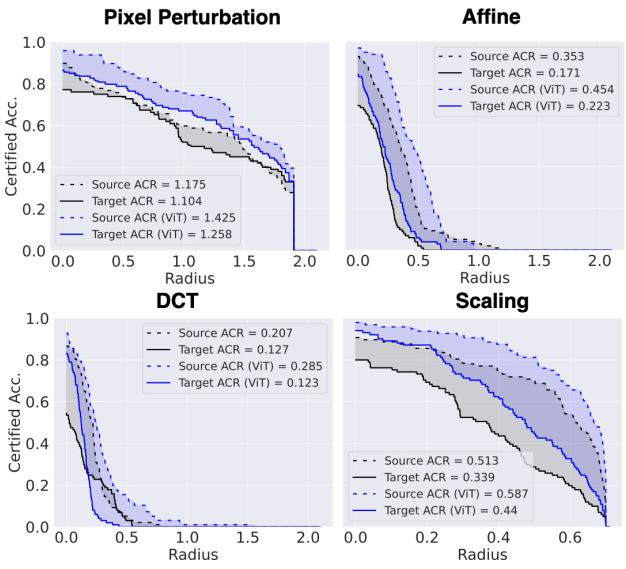


Figure 2: **Generalizability of Certified Robustness.** We certify ResNet-50 and ViT-Base against pixel perturbations and input deformations in the source and target domains of PACS. We observe that 1) certified robustness deployed via randomized smoothing generalizes to unseen domains, and that 2) a stronger architecture (ViT-Base) leads to a better source and target certified accuracy.

stronger backbone architecture can boost the CA generalizability. Finally, we evaluate how well perceptual similarity, as measured by FID and R-FID (Heusel et al. 2017; Alfarra et al. 2022c), predicts the generalization of certified robustness.

Q4: Can certified robustness, obtained via randomized smoothing, generalize to unseen domains? We train smooth classifiers on a collection of source domains and certify the models on both the source and target domains. The target domains are *unseen* before certification. We plot the source CA curve with dashed black lines and the target CA curve with solid blue in Figure 2, along with the corresponding ACR. Our results show that ❹ **a considerable portion of the certified robustness, acquired by randomized smoothing, is maintained in the unseen domain.** When certified against pixel perturbations in the unseen domain, the average certified radius of ResNet-50 drops by around 6% only. Utilizing DeformRS, we extend this result from simple pixel perturbations to geometric deformations, like scaling and affine transformations. This experiment is promising, since the models are never trained on the target data, but still exhibit some certified robustness. This validates the importance of recent research efforts that improve on randomized smoothing (Zhai et al. 2020; Alfarra et al. 2022b). *To address real-world security challenges, we encourage future certified robustness works to conduct experiments on domain generalization datasets.*

Q5: Does the target certified accuracy improve when the feature extractor is improved? To investigate the influence of the backbone architecture on the certified robustness of a deployed model, we change the architecture from ResNet-50 to ViT-Base and plot the target CA curve for ViT-Base in solid blue in Figure 2. We observe that the target ACR obtained by ViT-Base on PACS is higher than the target ACR obtained by ResNet-50 across deformations. ❺ **A significant improvement of the target certified robustness is achieved by using a stronger backbone architecture.** This result is consistent with the robustness literature (Gowal et al. 2020), where stronger backbones tend to exhibit better robustness, and the domain generalization literature (Gulrajani and Lopez-Paz 2021), where stronger backbones tend to exhibit better generalization accuracy. *We believe that research on models with better generalization can lead to better certified robustness in unseen domains.*

Q6: Does the generalizability of certified robustness correlate with the perceptual similarity between the source and target domains? In all previous experiments, we considered the average certified accuracy over all possible target domains. We now conduct a more fine-grained study to these target domains individually. We measure the drop in the average certified radius (ΔACR) between the source and target domains with the perceptual similarity between both domains captured by FID (Heusel et al. 2017) and the more robust R-FID (Alfarra et al. 2022c). To that end, we conduct experiments on PACS where we select one domain as the unseen target and treat the rest as source domains. We train a classifier on the source data and plot the certified accuracy curves against scaling and translation deformations on both the source and target domains in Figure 3 accompanied by ΔACR. We also report the FID and R-FID between the source and target domains. Note that *higher* FID/R-FID indicates *less* similarity of distributions. ❻ **Perceptual similarity, as captured by FID and R-FID, is not predictive of performance and robustness generalizability.** Surprisingly, the photo domain, which has the highest FID (34.3) and R-FID (87.7) scores, exhibits the largest certified accuracy generalization. In this case, the ACR for the target domain is higher than the source domain resulting in a negative ΔACR (-0.1 when certifying against translation). The appendix includes experiments with other deformations where we observe similar behavior. *We regard the development of a suitable distribution similarity metric, which better correlates with the level of generalizability, as an important research direction.*

6 Real-world Application: Medical Images

To demonstrate the applicability of the DG setup to real-world settings, we investigate the generalization of robustness in medical diagnostics. Data collected by medical imaging techniques, like computed tomography (CT) and magnetic resonance imaging (MRI), is susceptible to noise. This noise includes intensity variations caused by subject movement (Shaw et al. 2019), respiratory motion (Axel et al. 1986), quantum noise associated with X-rays (Hsieh 1998), and inhomogeneity in the MRI magnetic field (Leemput et al. 1999). Moreover, due to privacy concerns, a model trained in one medical institution should be deployed in another with limited data sharing and retraining (Kaissis et al.

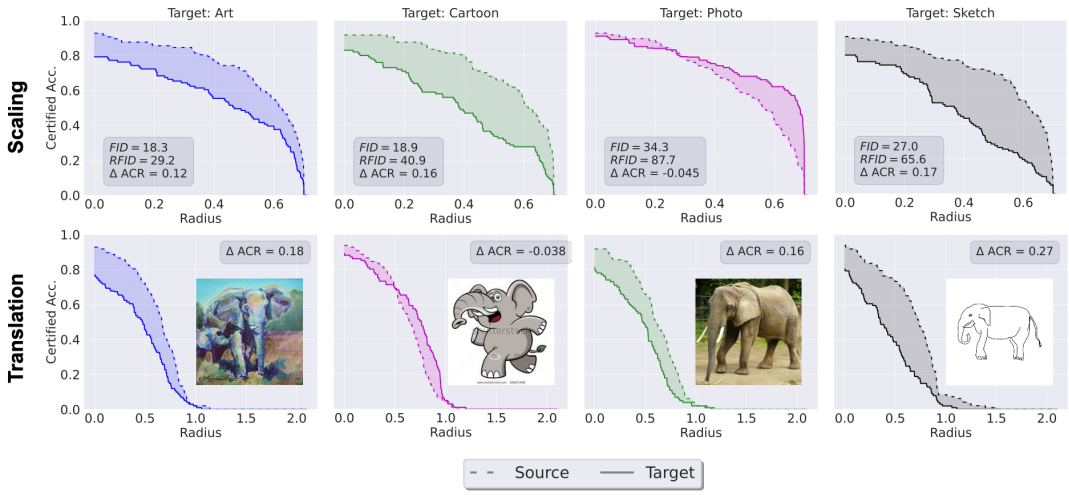
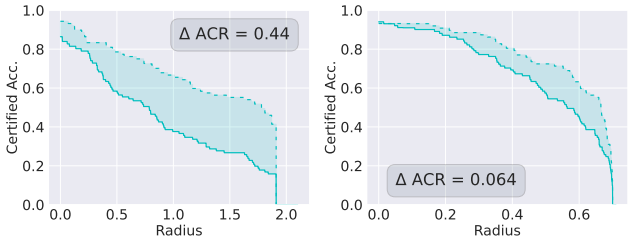


Figure 3: **Does visual similarity correlate with robustness generalizability?** We vary the target domain and plot the certified accuracy curves for two deformations: scaling and translation. A sample from each domain is shown in the second row. The FID/R-FID distances between the source domains and each target are reported in the first row. Visual similarity, measured by FID and R-FID, does not correlate with the level of robustness generalization to the target domain.



(a) Pixel Perturbations (b) Scaling

Figure 4: **Certified Robustness in the Medical Domain.** The generalization of robustness against pixel variations in medical images is critical. Yet, there is a gap in certified robustness when the DNN is deployed in an unseen hospital.

2020; Ziller et al. 2021; Liu et al. 2021). To test the generalization of robustness in this practical setup, we use the DG dataset WILDS CAMELYON17 (Bándi et al. 2019; Koh et al. 2021) to train models on tissue images from four hospitals and evaluate them on images from an unseen hospital. For the first time, in addition to domain generalization, we explore robustness in CAMELYON17. The task in WILDS CAMELYON17 is to predict whether a tissue image contains a cancerous tumor or not.

Adversarial Augmentation for Better Generalizability. We test the generalization accuracy of a standard-trained ($\epsilon = 0$) and a robust ($\epsilon = 8/255$) model by following setup from Section 4. In contrast to the domains studied in Table 2, adversarial training improves generalization to the unseen hospital. While the clean accuracy of the standard model is 94.05%, the clean accuracy of the robust model is 95.28%. This value is even competitive with the target accuracy (95.25%) obtained by the popular DG strategy CORAL (Sun and Saenko 2016). The robust accuracy also improves from 82.03% to 92.67%. This significant boost in domain generalization can be attributed to the similarity between pixel perturbations and the underlying domain shift in the medical images. *We encourage future works to study different adversarial training methods that go beyond pixel perturbations, and to propose application-specific augmentations for different distribution shifts.*

Certified Robustness. Next, we investigate the generalizability of certified robustness to the unseen hospital. We follow the experimental setup in Section 5 and measure the certified accuracy on the source and target domains. We observe from Figure 4 that some of the certified robustness generalizes to the unseen hospital when evaluated with pixel perturbations and scaling deformations. We include the results for other deformations in the appendix. We note that the drop in certified accuracy to the unseen hospital (given pixel perturbations) is 4 times what we saw in the PACS dataset in Section 5. This is concerning, as many sources of noise affect medical imaging data, so robust medical diagnostics is important for real-world adoption of AI for Health. *We encourage future research to develop better methods to close the target-source gap in certified robustness.*

7 Conclusion

We conducted a large scale empirical analysis to study the interplay between adversarial robustness and domain generalization. We deployed adversarial training and randomized smoothing as empirical and certified defenses. We found that both empirical and certified robustness generalize to unseen domains. We further included experiments on a real-world application, where adversarial training benefits both clean and robust accuracy in an unseen domain. Based on our findings, we encourage more research to understand: (i) under which conditions robust training improves the generalization accuracy, and (ii) what methods can improve certified accuracy in unseen domains.

References

Alfarra, M.; Bibi, A.; Khan, N.; Torr, P.; and Ghanem, B. 2022a. DeformRS: Certifying input deformations with randomized smoothing. In *Proc. of AAAI Conference on Artificial Intelligence*.

Alfarra, M.; Bibi, A.; Torr, P. H. S.; and Ghanem, B. 2022b. Data dependent randomized smoothing. In Cussens, J.; and Zhang, K., eds., *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 64–74. PMLR.

Alfarra, M.; Pérez, J. C.; Frühstück, A.; Torr, P. H.; Wonka, P.; and Ghanem, B. 2022c. On the Robustness of Quality Measures for GANs. *arXiv preprint arXiv:2201.13019*.

Axel, L.; Summers, R. M.; Kressel, H. Y.; and Charles, C. 1986. Respiratory effects in two-dimensional Fourier transform MR imaging. *Radiology*, 160.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Bándi, P.; Geessink, O.; Manson, Q.; Dijk, M. V.; Balkenhol, M.; Hermsen, M.; Bejnordi, B. E.; Lee, B.; Paeng, K.; Zhong, A.; Li, Q.; Zanjani, F. G.; Zinger, S.; Fukuta, K.; Komura, D.; Ovtcharov, V.; Cheng, S.; Zeng, S.; Thagaard, J.; Dahl, A. B.; Lin, H.; Chen, H.; Jacobsson, L.; Hedlund, M.; Çetin, M.; Halici, E.; Jackson, H.; Chen, R.; Both, F.; Franke, J.; Kusters-Vandevelde, H.; Vreuls, W.; Bult, P.; Ginneken, B. V.; Laak, J. V. D.; and Litjens, G. 2019. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, 38.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain Generalization by Solving Jigsaw Puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cha, J.; Chun, S.; Lee, K.; Cho, H.-C.; Park, S.; Lee, Y.; and Park, S. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 1310–1320. PMLR.

Delange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, Z.; Zhang, L.; Vodrahalli, K.; Kawaguchi, K.; and Zou, J. 2021. Adversarial Training Helps Transfer Learning via Better Representations. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; and Madry, A. 2019. Exploring the Landscape of Spatial Robustness. In *ICML*.

Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Gong, R.; Li, W.; Chen, Y.; and Gool, L. V. 2019. DLOW: Domain Flow for Adaptation and Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

Gowal, S.; Qin, C.; Uesato, J.; Mann, T.; and Kohli, P. 2020. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples.

Gu, S. S.; and Rigazio, L. 2015. Towards Deep Neural Network Architectures Robust to Adversarial Examples. *CoRR*, abs/1412.5068.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*.

Hsieh, J. 1998. Adaptive streak artifact reduction in computed tomography resulting from excessive x-ray photon noise. *Medical Physics*, 25.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B.; Haque, I.; Beery, S. M.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5637–5664. PMLR.

Lécuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D. J.; and Jana, S. S. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, 656–672.

Lee, S.; Lee, W.; Park, J.; and Lee, J. 2021. Towards Better Understanding of Training Certifiably Robust Models against Adversarial Examples. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Leemput, K. V.; Maes, F.; Vandermeulen, D.; and Suetens, P. 1999. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18.

Li, D.; Yang, Y.; Song, Y. Z.; and Hospedales, T. M. 2017. Deeper, Broader and Artier Domain Generalization. volume 2017-October.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021. FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1013–1023.

Lu, W.; Wang, J.; Li, H.; Chen, Y.; and Xie, X. 2022. Domain-invariant Feature Exploration for Domain Generalization. *Transactions on Machine Learning Research*.

Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; and Lu, F. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110: 107332.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mohapatra, J.; Weng, T.-W.; Chen, P.-Y.; Liu, S.; and Daniel, L. 2020. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 244–252.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified Deep Supervised Domain Adaptation and Generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Nguyen, A. T.; Tran, T.; Gal, Y.; and Baydin, A. G. 2021. Domain Invariant Representation Learning with Domain Density Transformations. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The Limitations of Deep Learning in Adversarial Settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Certified Defenses against Adversarial Examples. In *International Conference on Learning Representations*.

S., G. P.; Pérez, J. C.; Alfarra, M.; Giancola, S.; and Ghanem, B. 2022. 3DeformRS: Certifying Spatial Deformations on Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15169–15179.

Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; and Madry, A. 2020. Do Adversarially Robust ImageNet Models Transfer Better? In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3533–3545. Curran Associates, Inc.

Shafahi, A.; Saadatpanah, P.; Zhu, C.; Ghiasi, A.; Studer, C.; Jacobs, D.; and Goldstein, T. 2020. Adversarially robust transfer learning. In *International Conference on Learning Representations*.

Shaw, R.; Sudre, C.; Ourselin, S.; and Cardoso, M. J. 2019. MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty. In Cardoso, M. J.; Feragen, A.; Glocker, B.; Konukoglu, E.; Oguz, I.; Unal, G.; and Vercauteren, T., eds., *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, 427–436. PMLR.

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards Out-Of-Distribution Generalization: A Survey.

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. arXiv:1607.01719.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. arXiv:1412.3474.

Utrera, F.; Kravitz, E.; Erichson, N. B.; Khanna, R.; and Mahoney, M. W. 2021. Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification. In *International Conference on Learning Representations*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; and Qin, T. 2021. Generalizing to Unseen Domains: A Survey on Domain Generalization. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4627–4635. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wang, M.; and Deng, W. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *International Conference on Learning Representations*.

Wu, D.; Xia, S.-T.; and Wang, Y. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 2958–2969. Curran Associates, Inc.

Zhai, R.; Dan, C.; He, D.; Zhang, H.; Gong, B.; Ravikumar, P.; Hsieh, C.-J.; and Wang, L. 2020. MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius. In *International Conference on Learning Representations (ICLR)*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L. E.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. volume 97, 7472–7482. PMLR.

Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; and Shen, Z. 2021. Deep Stable Learning for Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5372–5382.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578. Springer.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.

Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; and Kaissis, G. 2021. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11.