Certified Causal Defense with Generalizable Robustness

Yiran Qiao[1], Yu Yin[1], Chen Chen[2], Jing Ma[1*]

[1]Case Wester Reserve University
[2]University of Virginia
yxq350@case.edu, yxy1421@case.edu, zrh6du@virginia.edu, jing.ma5@case.edu

Abstract

While machine learning models have proven effective across various scenarios, it is widely acknowledged that many models are vulnerable to adversarial attacks. Recently, numerous efforts have emerged in adversarial defense. Among them, certified defense is well known for its theoretical guarantees against arbitrary adversarial perturbations on input within a certain range. However, most existing works in this line struggle to generalize their certified robustness in other data domains with distribution shifts. This issue is rooted in the difficulty of eliminating the negative impact of spurious correlations on robustness in different domains. To address this problem, in this work, we propose a novel certified defense framework GLEAN, which incorporates a causal perspective into the generalization problem in certified defense. More specifically, our framework integrates a certifiable causal factor learning component to disentangle the causal relations and spurious correlations between input and label, thereby excluding the negative effect of spurious correlations on defense. On top of that, we design a causally certified defense strategy to handle adversarial attacks on latent causal factors. In this way, our framework is not only robust against malicious noises on data in the training distribution but also can generalize its robustness across domains with distribution shifts. Extensive experiments on benchmark datasets validate the superiority of our framework in certified robustness generalization in different data domains.

Extended version — https://arxiv.org/pdf/2408.15451

Introduction

Machine learning (ML) models, particularly deep neural networks (DNN), have demonstrated remarkable success across a diverse range of areas (Devlin et al. 2018; Silver et al. 2017; He et al. 2016). Despite their success, these models still exhibit significant vulnerabilities to adversarial perturbations on input (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014; Biggio et al. 2013). A typical example in image classification is that a trained classifier that correctly classifies an image x can be easily fooled by a perturbed image $x+\delta$, where δ represents adversarial perturbations that are imperceptible to human perception. This weakness impedes the deployment of ML models in critical applications where security and reliability are priorities, such as autonomous driving and healthcare.

In the past few decades, researchers have developed numerous defense methods to enhance the adversarial robustness of ML models. Many of these methods are based on *adversarial training* (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017; Zhang et al. 2019; Athalye, Carlini, and Wagner 2018), which incorporates adversarial samples into model training. Despite its impressive performance, adversarial training is an empirical approach that lacks theoretical guarantees. That is, although it can enhance robustness against certain types of attacks, it may still be vulnerable to other unknown, or more potent adversarial perturbations. Differently, another line of work develops *certified robustness*. A certified robust classifier can theoretically guarantee that its prediction for a point x remains constant within a certain specified range (a.k.a. *radius*) of perturbations on x, regardless of the type of attack. Randomized smoothing-based certified defense (Lecuyer et al. 2019; Li et al. 2018; Cohen, Rosenfeld, and Kolter 2019) is one of the most representative methods in this area. Specifically, given an arbitrary base classifier f, this method can convert it to a certifiably robust classifier g, which is created by randomly sampling multiple noised versions of a given input and using the aggregated output from these variations to make final predictions. Inspired by this approach, many subsequent studies (Li et al. 2019; Jeong and Shin 2020; Jeong et al. 2021; Salman et al. 2019; Zhai et al. 2020) have expanded upon the basis of random smoothing.

Currently, most existing certified defense works focus on data in the same domain yet overlook other domains with distribution shifts. This limitation can result in a markedly degraded certified robustness performance when these methods are applied to the test domain (Sun et al. 2021). As discussed in previous work (Ilyas et al. 2019; Beery, Van Horn, and Perona 2018), such degradation of robustness lies in the fact that ML models tend to overfit spurious correlations between features and labels. As these spurious correlations often vary across different domains (Ye et al. 2022), fitting spurious correlations can easily lead g to make incorrect predictions or correct predictions but with lower confidence levels. The former results in the certified radius being assigned zero, while the latter also leads to a reduced certified radius. Therefore, domain shifts can lead to weak generalization w.r.t. not only prediction performance but also certified robustness.

Different from ML models, humans can naturally capture the invariant relations between labels and their corresponding causal factors, various studies (Zhang, Zhang, and Li 2020; Tang, Huang, and Zhang 2020; Schölkopf et al. 2021) argue that human's inherent causal view brings a solution to avoid the domain generalization hurdle for robustness. Inspired by this, in this paper, we study the problem of *generalizing the certified robustness under domain shifts from a causal view*.

However, addressing this problem presents multifaceted challenges. **Challenge 1:** As aforementioned, spurious correlations varying across domains adversely affect robustness. To achieve robustness across domains, it is crucial to effectively remove the impact of these spurious correlations. However, identifying and eliminating the impact of spurious correlations on robustness in different domains presents a challenge. **Challenge 2:** Apart from spurious factors, the distribution shifts often bring challenges for the model to defend against the perturbations on the factors that causally determine the label in unseen domains, which leads to diminished certified robustness. **Challenge 3:** It is important to provide theoretical guarantees for robustness on other data domains, but most existing works remain in empirical observations and lack theoretical analysis. Although some works (Salman et al. 2019; Zhai et al. 2020) in certified defenses has provided upper bounds on perturbations while maintaining robustness, they were not designed to address certified robustness in the domain shift context.

In this work, to tackle these challenges, we propose a novel framework **G**enera**L**izable c**E**rtified c**A**usal defe**N**se (**GLEAN**) that enhances the certified robustness of models on data in different domains. To mitigate the influence of spurious correlations on robustness generalization (Challenge 1), we construct a causal model for data in different domains, and simultaneously conduct a causal analysis on model robustness and generalization. Based on the causal model, we filter out the impact from spurious correlations and enhance robustness across domains. This is different from most existing defense algorithms which take the same strategy indiscriminately towards all the input features. To achieve certified robustness through causal factors (Challenge 2), we utilize a certified causal factor learning module with Lipschitz constraint. This module enables certification through the latent representation space for high-level causal factors, conducting certified defense for perturbations on causal factors that determine the label in different domains. To bring theoretical guarantees for robustness on different data domains (Challenge 3), we derive a theoretical analysis by leveraging the theoretical support of certified defense and causal inference. Our main contributions can be summarized as:

- We investigate an important but underexplored problem of certified defense on data in different domains. We analyze the significance of this research problem and its corresponding challenges.
- We propose a novel causality-inspired framework GLEAN for this problem, extending certified robustness across data domains. Specifically, we develop a certified defense strategy based on certifiable causal factor learning, which excludes spurious correlations and provides a certified radius for test data with a theoretical guarantee.

- We conduct extensive experiments to evaluate our framework on both synthetic and real-world datasets. The results show that our framework significantly outperforms the prevalent baseline methods.

Preliminaries and Related Work

We consider a classification task with x representing an input instance and y denoting the corresponding label, where $x \in \mathcal{X}, y \in \mathcal{Y} := \{1, ..., K\}$, $\mathcal{X}$ and $\mathcal{Y}$ represent the input space and the label space, respectively. A classifier trained for this task can be denoted by $f : \mathcal{X} \to \mathcal{Y}$. The data may be collected from different domains (i.e., environments). We use the superscript $(\cdot)^d$ to denote the data in a certain domain d.

Certified Defense

Robust Radius The robust radius for an instance x is the largest range (e.g., a l_2 ball) centered at x, within which f provides a correct prediction for x, and this prediction remains constant. It is defined as follows:

$$R(f; x, y) = \min_{f(x') \neq f(x)} \|x' - x\|_2. \tag{1}$$

Unfortunately, calculating the robust radius for neural networks is proven to be an NP-complete problem (Katz et al. 2017; Sinha et al. 2017) thus both challenging and time-consuming.

Certified Radius Many previous works proposed certification methods to derive a *certified radius* that is the lower bound of the robust radius. Research in this area falls into two categories: exact methods and conservative methods. Exact methods (Ehlers 2017; Bunel et al. 2018; Tjeng, Xiao, and Tedrake 2017), usually based on Satisfiability Modulo Theories or mixed integer linear programming, guarantee the identification of a perturbation δ within a radius r that can cause f to change its prediction. However, they require the model to have a limited scale. Conservative methods (Wong and Kolter 2018; Wong et al. 2018; Gowal et al. 2018) ensure the detection of existing adversarial examples and, in addition, refuse to make certification for some vulnerable data points. These methods, though more scalable, impose specific assumptions on the model's architecture.

Randomized Smoothing Randomized Smoothing (RS) is proposed to tackle the above limitations, which can be applied to any architectures (Cohen, Rosenfeld, and Kolter 2019). It constructs a smoothed classifier g from an arbitrary base classifier f. The definition of g is as follows:

$$g(x) = \arg\max_{y \in \mathcal{Y}} P(f(x + \eta) = y) \tag{2}$$

In this formula, $\eta \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is the isotropic Gaussian noise with the noise level σ as the hyperparameter. The smoothed classifier g can be summarized as returning the class most likely to be predicted by f when the input x is sampled from Gaussian distributions. (Cohen, Rosenfeld, and Kolter 2019) provide the theoretical form of certified radius which is the lower bound of the robust radius:

$$CR(f; x, y) = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)), \tag{3}$$

where $p_A = P(f(x + \eta) = y_A)$, $p_B = \max_{y \neq y_A} P(f(x + \eta) = y)$, meaning that f will mostly return class y_A with the
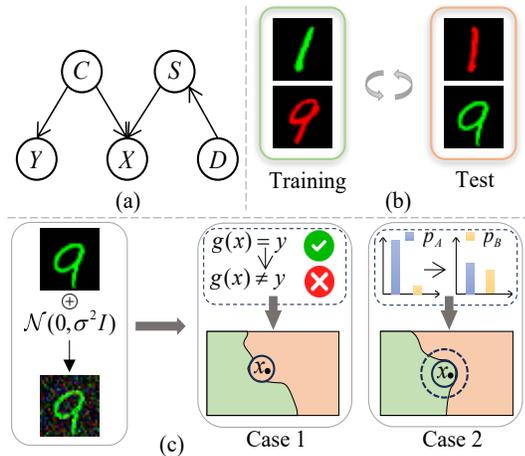
Figure 1: (a) Causal graph of data generation across domains; (b) A showcase of domain shift, here we use images in CMNIST as an example. (c) Two common cases of domain shifts leading to decreased ACR in certification. The pink area represents an incorrect decision area, green signifies the correct decision area, and circles represent a robust l_2 ball.

probability p_A, and will return the runner-up class with the probability p_B. Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function. Then $g(x + \delta) = y_A$ for all $||\delta||_2 \leq CR$.

Reduced Certified Robustness in Unseen Domains

There widely exist distribution shifts between data in different domains, i.e., $P^d(X, Y) \neq P^{d'}(X, Y)$, where $d \neq d'$. Inspired by (Zhang et al. 2013; Pearl, Glymour, and Jewell 2016), we use the causal graph shown in Figure 1 (a) to illustrate the causal model for data across domains. Specifically, as shown in Figure 1, we discuss the causal relations among five variables: label Y (e.g., an object type), input features X (e.g., an image), causal factors C (e.g., the object shape in the image) that determine the label, non-causal factors S (e.g., the background in the image), and domain variable D (e.g., the data source). C and S are usually high-level latent concepts without observed supervision. Noticeably, S often has spurious correlations with Y, even if they are not causally related. Such spurious correlations often vary in different domains, i.e., $P^d(Y|S) \neq P^{d'}(Y|S)$.

Distribution shift often brings challenges in certified robustness (Sun et al. 2022). Here, we use a simple experiment to show the rapid deterioration of certified robustness on data in different domains, where the task is to classify the Colored-MNIST (CMNIST) dataset (Arjovsky et al. 2019). CMNIST is a modified version of the handwritten digit image dataset MNIST (LeCun et al. 1998), artificially constructed to include two colors, red and green. The colors are strongly but spuriously correlated with the label Y ($Y = 0$ for digits $0 \sim 4$, and $Y = 1$ for digits $5 \sim 9$). For this dataset, the color of the digits is a non-causal factor S, while the shape of the digits is the causal factor C. The spurious correlation between Y and S in the training domain

is reversed in the test domain, as shown in Figure 1 (b).

In the CMNIST dataset, it is unsurprising that a classifier relying on the digit color would fail on the test domain due to the shift in the spurious correlation $P(Y|S)$. To show the negative impact of spurious correlation on certified robustness, we compare the Average Certified Radius (ACR, a metric used to evaluate certified robustness) of random smoothing-based certified defense on CMNIST with the results on MNIST (where there is no digit color and thus the above spurious correlations do not exist). As observed from the results in Table 1, there is a significant degradation of prediction accuracy and ACR on the test domain, indicating severe issues for certified defense under domain shift.

L-Lipschitz Networks

Definition 1 (Lipschitz Continuity). *A function $f : X \rightarrow Y$ is called Lipschitz continuous if there exists a non-negative constant L (known as the Lipschitz constant) such that for all $x_1, x_2 \in X$ the following condition is met:*

$$||f(x_1) - f(x_2)||_2 \leq L||x_1 - x_2||_2. \tag{4}$$

Based on the definition, if a neural network f is 1-Lipschitz, then for any input x, the output y satisfies $||\mathbf{y}||_2 \leq ||\mathbf{x}||_2$. Equivalently, $||\mathbf{y_1} - \mathbf{y_2}||_2 \leq ||\mathbf{x_1} - \mathbf{x_2}||_2$.

GLEAN: Framework and Theories

In this section, we introduce the detailed technologies and theories in our proposed framework. We begin by proposing a causal view of robustness under domain shifts. Next, we introduce our design of a certifiable causal factor learning module to exclude the impact of spurious correlations on robustness. Then, we explain the whole certified defense process through the latent causal space, providing a theoretical guarantee for the certified robustness of data in different domains.

Causal View of Robustness and Cross-Domain Generalization

As introduced in the last section, we illustrate our causal graph in Figure 1 (a). Noticeably, although the spurious correlations vary in different domains, since $C \rightarrow Y$ has a direct causal link, the relationship between C and Y remains invariant across domains and is thus unaffected by domains. This inspired invariant learning based on the following causal invariance assumption (Li et al. 2022):

Assumption 1 (Causal Invariance over Domain Shifts) *For any two domains d and d', the probability $P(Y|C)$ is invariant to domain shifts, i.e.,:*

$$P^d(Y|C) = P^{d'}(Y|C), \forall d, d' \in \mathbb{D}, \tag{5}$$

Dataset	Test Acc	ACR ($\sigma = 0.25$)
CMNIST	21.01%	0.07
MNIST	72.03%	0.37

Table 1: Comparison of the certified defense performance with/without domain shift. The metrics include the prediction accuracy and the Average Certified Radius (ACR).

where $\mathbb{D}$ is the set of all possible domains.

Based on this assumption, a model that can identify causal factors and make predictions based on them can be generalized to unseen domains.

From a robustness perspective, distribution shifts introduce significant additional challenges. At a high level, robustness can be viewed as a generalization problem over an adversarial distribution (Xin et al. 2023). This adversarial distribution often differs from the unseen domains derived from natural distributions, necessitating more sophisticated methods to capture high-level causal factors in decision-making while filtering out the impact of adversarial perturbations. More specifically, for a target domain d', we have:

$$P^{d'}(Y|X) = \int_{c \in \mathbb{C}} P^{d'}(c|X)P^{d'}(Y|c) = \int_{c \in \mathbb{C}} P^{d'}(c|X)P^d(Y|c), \quad (6)$$

$$P^{d'}(Y|X) = \int_{s \in \mathbb{S}} P^{d'}(s|X)P^{d'}(Y|X,s), \quad (7)$$

where $\mathbb{C}$ and $\mathbb{S}$ are the space of C and S, respectively. Here, $d \neq d'$. Each of the equations above decomposes $P(Y|X)$ into two components. As shown in Eq. 6, the model can generalize to a different (or even adversarial) domain d' if it accurately captures the causal factors C from X. The other term $P(Y|C)$ remains invariant across domains, which helps to mitigate the risk of increased vulnerability in new domains. However, as indicated by Eq. (7), $P^{d'}(Y|X,s)$ varies across domains, which can increase the vulnerability to adversarial perturbations. This increased vulnerability stems from two main issues, as illustrated in Figure 1 (c): 1) reduced accuracy in the test domain, leading to diminished prediction reliability even under slight perturbations; and 2) the change in $P^{d'}(Y|X,s)$ across domains increases decision uncertainty at each $X = x$ due to potential conflicts between $P(Y|C)$ and $P(Y|S)$. These factors collectively complicate the task of achieving robustness across different domains. The above analysis indicates the importance of incorporating a causal view into the robustness problem across domains. In our framework, we identify the causal factors from input (i.e., modeling $P(C|X)$) with certifiable robustness, and conduct certified defense based on an invariant predictor $P(Y|C)$.

Causal Encoder with Lipschitz Constraint

Inspired by the above analysis and the observations of the aforementioned toy experiment, to achieve robustness in different domains, we develop a method to robustly identify causal factors from the input for downstream prediction. It is worth mentioning that, for many real-world scenarios, identifying causal factors in the input space (e.g., image pixels) is difficult without segmentation labels, and also less meaningful, because causal factors are often high-level concepts. Therefore, our method is built upon a representation space, where we conduct two main tasks: (1) learn the causal factors from the input features with an encoder $\Psi(\cdot)$; (2) provide a certifiable guarantee for robustness in this process.

For the first task, encouraged by recent progress in causal generalization, we extract the causal factors of input features in the latent space through techniques in invariant learning

(Krueger et al. 2021; Ahuja et al. 2020; Arjovsky et al. 2019; Mitrovic et al. 2020), which capture invariant factors across different domains. We can adopt one of the cutting-edge methods of this type for our invariant learning module. In this work, we leverage one of the most representative methods: invariant risk minimization (IRM) (Arjovsky et al. 2019) with the following optimization loss:

$$\mathcal{L}_{\mathrm{IRM}} = \sum_{d \in \mathbb{D}_{\mathrm{tr}}} R^d(\beta \circ \Psi) + \lambda \cdot \|\nabla_{w|w=1.0} R^d(w(\beta \circ \Psi))\|^2, \quad (8)$$

where $R^d(\beta \circ \Psi) = \mathbb{E}[\mathcal{L}(g(\Psi(x)), y)]$ is the prediction loss in domain d with an encoder Ψ and classifier β, w is a "dummy" classifier and can be fixed as a scalar 1.0, and $\mathbb{D}_{\mathrm{tr}}$ is the set that includes all training domains. According to (Arjovsky et al. 2019), the gradient of $R^d(w(\beta \circ \Psi))$ reflects the invariance of the learned latent representations. The non-negative hyperparameter λ controls the balance between the predictive ability and invariance.

Even though causal factor learning usually does not have specific restrictions regarding the encoder architecture, it is worth noting that an arbitrary architecture cannot provide certifiable robustness in the latent space. Therefore, for the second task, we adopt the 1-Lipschitz network (Trockman and Kolter 2021) to derive certifiable robustness across domains.

Certified Robustness for Unseen Domains

While significant progress has been made in certified defenses when training and test data share the same distribution, there is still limited exploration and a lack of theoretical guarantees for certified robustness under domain shifts. In this subsection, we bridge this gap by utilizing the theoretical support from certified defense (Cohen, Rosenfeld, and Kolter 2019) and causal inference to derive necessary theorems in this setting.

According to previous discussions, we perform random smoothing for the causal factors in the latent space. Therefore, based on the calculation of the certified radius, we introduce the following Theorem 1:

Theorem 1. *Suppose we have a causal encoder $\Psi : \mathcal{X} \rightarrow \mathcal{Z}$, and an arbitrary classifier $\beta : \mathcal{Z} \rightarrow \mathcal{Y}$. Let g_β be defined as $g_\beta(z) = \arg\max_{y \in \mathcal{Y}} P(\beta(z + \eta) = y)$, where $\eta \sim \mathcal{N}(0, \sigma^2 I)$, $z = \Psi(x)$ is the latent causal representation. Suppose $\underline{p_A}$ is the lower bound of p_A, $\overline{p_B}$ is the upper bound of p_B (here $\underline{p_A}$ and $\overline{p_B}$ are obtained based on g_β with input in the representation space), $\underline{p_A}, \overline{p_B} \in [0, 1]$ and satisfy:*

$$P(\beta(z + \eta) = y_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{y \neq y_A} P(\beta(z + \eta) = y). \quad (9)$$

Then $\forall d, d' \in \mathbb{D}$, $g_\beta(z^d + \delta_z) = g_\beta(z^{d'} + \delta_z) = y_A$ for all $\|\delta_z\|_2 < CR_z$, where δ_z is the perturbation applied to the latent causal representation z and

$$CR_z(\beta; x, y) = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})). \quad (10)$$

This theorem provides us a theoretical guarantee that any perturbation δ_z within the range CR_z will not change the prediction of the smoothed classifier g_β. It also provides a theoretical guarantee for generalization: for any two instances from d
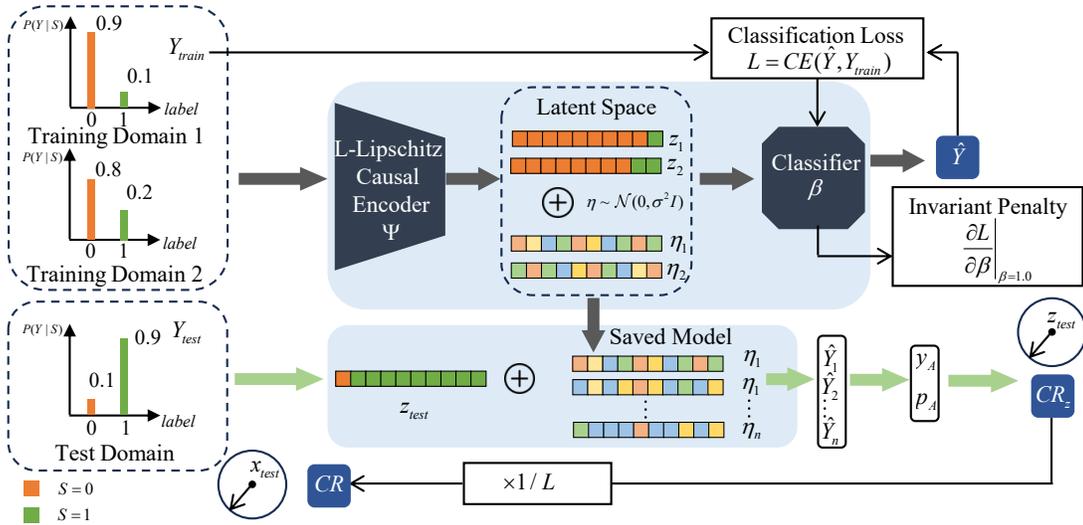
Figure 2: An overview of the proposed framework GLEAN. The upper part represents the training process, while the lower part depicts the certification process on the test domain. Here, we showcase two training domains and one test domain with two classes 0 and 1, where the color of the object is a spurious factor. We define $S = 0$ as orange and $S = 1$ as green. In training domain 1, the spurious distribution between color and category is $P(Y = 0|S = 0) = 0.9$ and $P(Y = 1|S = 1) = 0.1$. These values change to 0.8 and 0.2, respectively in training domain 2, and then to 0.1 and 0.9 in the test domain. Thus, there is a correlation shift between the different domains of this dataset. The causal encoder is equipped with Lipschitz constraints with Lipschitz constant L. Y_{train} and Y_{test} are ground truth labels. $\hat{Y}$ is the predicted label. z is the causal latent representations and each η is a Gaussian noise. y_A is the most probable class among all the $\hat{Y}$ after sampling with the probability p_A. Then we can leverage p_A to compute the certified radius in latent space CR_z and finally revert it back to get the certified radius CR in input space.

and d' respectively, if their causal latent representations (denoted by z^d and $z^{d'}$) learned from the causal encoder Ψ are the same, then the predictions of g_β for them are consistent. Moreover, the certified radius for z across these domains will also be consistent. Therefore, Theorem 1 provides theoretical support for achieving certified robustness on data in different domains by performing random smoothing in the latent space.

Another significant problem left is that the certified radius CR_z mentioned in Theorem 1 is obtained by applying Gaussian noise within latent space and then performing Monte Carlo sampling. Thus, the robustness guarantee is only for β. However, in practice, attackers often directly perturb input features. Therefore, the certified radius obtained in the latent space needs to be mapped back to the input space to provide certified robustness for the entire classifier f. Correspondingly, we have Theorem 2 as follows:

Theorem 2. *Let the causal encoder Ψ be L-Lipschitz. Let g be defined as $g(x) = \arg\max\limits_{y \in \mathcal{Y}} P(\beta(\Psi(x + \eta)) = y)$. Then*

$g(x^d + \delta) = g(x^{d'} + \delta) = y_A$ *for all $\|\delta\|_2 < CR_z/L$, where δ is the perturbation applied to input features x.*

Briefly, if we use an L-Lipschitz neural network in the causal factor learning module, we can calculate the certified radius in the input space. This is because we can simply scale the certified radius in the latent space by the Lipschitz constant L, such that $CR \geq CR_z/L$. If $L = 1$, then CR_z will be the lower bound of CR. With the aforementioned causal invariant assumption, the certified robustness for instances

in one domain can also be propagated to instances in other domains with the same causal factors. Therefore, we are able to provide theoretical guarantees for cross-domain certified robustness. Detailed proofs can be found in the Appendix.

Implementation

Overview of Framework We integrate the previous methods and theories to form our framework, which is demonstrated in Figure 2. In Figure 2, the gray path represents the training process. During training, we apply Gaussian augmentation to z to enhance the prediction accuracy during the RS phase. The green path represents the certifying process. We first train the causal encoder Ψ and classifier β, then obtain robustness guarantees for the classifier β by adding Gaussian noise to z with Monte Carlo sampling. The bottom path represents the mapping process. Specifically, it involves multiplying the certified radius in the latent space by the mapping constant $1/L$, and reverting back to the input space to obtain robustness guarantees for the input feature x.

Architecture As aforementioned, we use Lipschitz constraints in the causal factor learning module. We define the final linear layer as the classifier β, with all preceding layers forming the encoder Ψ. We apply the Cayley transform (Trockman and Kolter 2021) to achieve orthogonality, thereby ensuring that each linear layer has a Lipschitz constant of 1. For the activation functions, we employ GroupSort (Anil, Lucas, and Grosse 2019), which also has 1-Lipschitzness. More details on the implementation

Datasets	Models	$r = 0.00$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	ACR
	Gaussian	18.1	14.8	12.9	10.5	9.0	8.6	8.1	7.8	7.3	6.0	0.0458
	MACER	23.6	19.4	15.5	12.4	10.1	8.8	7.4	5.6	4.1	1.9	0.0482
CMNIST	SmoothAdv	27.2	22.7	16.9	13.6	10.5	8.5	7.3	5.8	3.6	1.3	0.0518
	Consistency	12.1	11.3	10.8	10.7	10.6	10.5	10.4	10.4	10.3	10.3	0.0488
	GLEAN (Ours)	**64.3**	**62.7**	**60.5**	**58.0**	**55.8**	**53.2**	**51.4**	**47.9**	**44.9**	**38.6**	**0.2466**
	Gaussian	31.0	31.0	31.0	29.0	28.0	26.0	26.0	24.0	22.0	17.0	0.1218
	MACER	21.0	19.0	15.0	13.0	10.0	10.0	10.0	8.0	7.0	6.0	0.053
CelebA	SmoothAdv	25.0	23.0	18.0	16.0	14.0	11.0	10.0	9.0	9.0	5.0	0.0623
	Consistency	24.0	24.0	23.0	23.0	22.0	21.0	20.0	20.0	20.0	20.0	0.0989
	GLEAN (Ours)	**62.0**	**59.0**	**57.0**	**56.0**	**52.0**	**51.0**	**49.0**	**45.0**	**42.0**	**38.0**	**0.2326**
	Gaussian	33.0	32.0	32.0	31.0	28.0	25.0	24.0	23.0	22.0	19.0	0.1223
	MACER	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	27.0	27.0	0.1272
DomainNet	SmoothAdv	27.0	27.0	26.0	26.0	25.0	24.0	23.0	23.0	20.0	18.0	0.1097
	Consistency	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0	0.1234
	GLEAN (Ours)	**63.0**	**61.0**	**55.0**	**54.0**	**48.0**	**43.0**	**41.0**	**37.0**	**30.0**	**23.0**	**0.2088**

Table 2: A comparison of certified test accuracy (%) and ACR between our framework and baselines. For each method, we recorded data for ten radii r ranging from 0.00 to 0.45, with increments of 0.05. Every model is certified with $\sigma = 0.12$. We highlight our results in bold whenever the value improves the baselines.

of 1-Lipschitz networks can be found in the Appendix.

Experiments

In this section, we conduct extensive experiments to evaluate our framework on one synthetic dataset and two real-world datasets. Specifically, we answer the following research question based on the experimental results. **RQ1**: How does GLEAN perform compared to the baselines of certified defense? **RQ2**: How do different components in GLEAN contribute to the performance? **RQ3**: How does GLEAN perform under different settings of hyperparameters?

Datasets

We introduce the three datasets used in the experiments: CMNIST (Arjovsky et al. 2019), CelebA (Liu et al. 2015) and DomainNet (Peng et al. 2019). Detailed information on the domain construction and division of all these three datasets can be found in the Appendix.

Experiment Settings

Baselines. We evaluate our framework by comparing it against several representative certified defense methods. All these methods are based on RS:

- **Gaussian** (Cohen, Rosenfeld, and Kolter 2019): Standard training with Gaussian noise based random smoothing.
- **MACER** (Zhai et al. 2020): Add a regularization term that maximizes an approximate form of the certified radius.
- **SmoothAdv** (Salman et al. 2019): Adversarial training is incorporated during the training of the smoothed classifier.
- **Consistency** (Jeong and Shin 2020): The Kullback-Leibler divergence between the mean of the classifier's predictions after various perturbations and the prediction after a single perturbation was used as a regularization term. This term minimizes the variance in the classifier's predictions after different perturbations, optimizing the objective for robust training of the smoothed classifier.

Evaluation Metrics. We consider two widely-used evaluation metrics: (1) *certified accuracy* at different *radii*, which is defined as the fraction of the test set that CERTIFY (Cohen, Rosenfeld, and Kolter 2019) classifies correctly. CERTIFY is a practical Monte Carlo-based certification procedure that offers the prediction of g along with the lower bound of the certified radius or abstains the certification by sampling over n Gaussian noises with the probability of at least $1 - \alpha$, α is the significance level; (2) *average certified radius* (ACR), which is defined as $ACR = \frac{1}{|\mathbb{D}_{\texttt{test}}|} \sum_{(x,y) \in \mathbb{D}_{\texttt{test}}} CR(f; \sigma, x, y) \cdot \mathbf{1}_{[g(x,\sigma)=y]}$. Here, $|\mathbb{D}_{\texttt{test}}|$ is the capacity of the test set, CR is the certified radius returned by CERTIFY, $\mathbf{1}$ is the indicator function. We assign 0 to CR for incorrect prediction of g. We use the same settings in (Cohen, Rosenfeld, and Kolter 2019) with $n = 100000, n_0 = 100, \alpha = 0.001$ to apply CERTIFY. Here n_0 is the small number of samples to find y_A. Note that, for two different models, their certified accuracies sometimes cannot be directly compared. At a specific radius r, one model may have a higher certified accuracy than the other, but the situation may be reversed at another radius. Therefore, ACR is a more suitable metric as it reflects "average robustness".

Training Details. We use a three-layer MLP for CMNIST and a four-layer CNN for CelebA and DomainNet. During inference, we apply RS with the noise level $\sigma = 0.12$. The result of other σ is shown in the Appendix. We set the parameter of the regularization term $\lambda = 10000$ for all datasets.

Experiment Results

Performance For all datasets, more detailed settings for the training parameters are provided in the Appendix. Table 2 shows a comparison of performance between our framework and baselines w.r.t. ACR and the certified test accuracy with different radii r. We also plot the radius-certified accuracy curve in Figure 3. Note that ACR is equivalent to the area under the curve. From the results, we observe that our method achieves the highest certified accuracy and ACR (with a significant improvement compared with others) at almost all

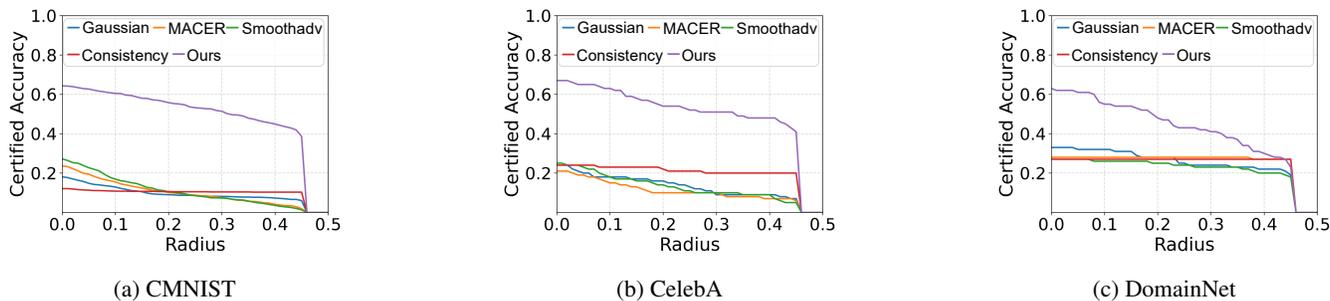(a) CMNIST   (b) CelebA   (c) DomainNet

Figure 3: Comparison of certified accuracy obtained using different methods across three datasets. The sharp decline at the end of the curves is due to a hard upper bound in the certification process for a variance σ and the number of Gaussian samples n.
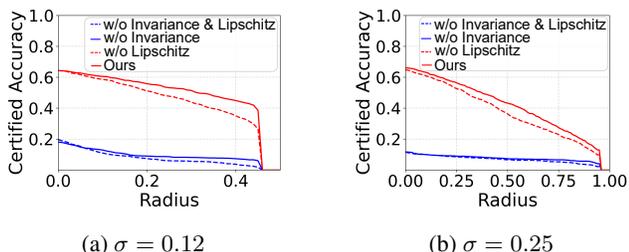

(a) $\sigma = 0.12$ (b) $\sigma = 0.25$

Figure 4: Ablation study on CMNIST with different σ.

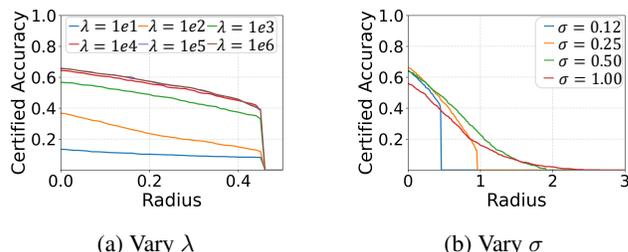(a) Vary λ (b) Vary σ

Figure 5: Performance of GLEAN under different parameters.

radii across the three datasets. Given that our training and testing data reside in different domains, the experimental results demonstrate that our approach significantly and consistently outperforms baselines in the generalization of certified robustness across domains. We omit the variance of the experimental results because it is far smaller than the performance gap between the methods. From Table 2, we can also observe that the ACR decreases progressively from CMNIST to CelebA to DomainNet. This decline is reasonable because the CMNIST dataset only involves spurious correlations between color and digits, whereas CelebA, in addition to the constructed spurious correlation between smiling and hair color, includes more complicated domain shifts regarding other facial features. For DomainNet, the complex variation in backgrounds makes the causal relationships within the data more difficult to capture.

Ablation Study To evaluate the effectiveness of each component in our method, we provide ablation study with the following variants: (1) w/o invariance: We remove the invariance regularization term in Eq.(8) and only use the first term as an ERM loss. (2) Network without Lipschitz Constraints: We replace the 1-Lipschitz layers in the network with the ones without any constraints. We conducted comparisons with two types of ablation studies simultaneously under $\sigma = 0.12$ and $\sigma = 0.25$. As shown in Figure 4, our model undoubtedly outperforms the version without the invariant penalty since this variant cannot capture the causal factors effectively and thus fails to mitigate the influence of spurious correlations on robustness. For our model, the certified accuracy and ACR with Lipschitz constraints is slightly better than that of networks without any constraints. This is because Lipschitz constraints ensure that we use causal factors for certification. The results

of the other two datasets are provided in the Appendix.

Parameter Study We set the hyperparameter $\lambda \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$, $\sigma \in \{0.25, 0.50, 1.00\}$. The results of the parameter study on CMNIST are shown in Figure 5. The results for the other two datasets are provided in the Appendix. We can observe in Figure 5 (a) that when λ increases, the certified accuracy at the same radius also increases, this is because a higher λ leads to stronger causal factor learning, and achieving stronger generalizable robustness. However, when λ exceeds 10,000, the improvement in model performance becomes negligible as it has reached the bottleneck of the model's ability to learn invariant causal factors. As shown in Figure 5 (b) σ controls the level of noise. A higher noise level means that we can obtain a larger certified radius but at the cost of reduced certified accuracy.

Conclusion

In this paper, we address the critical problem of generalizing certified robustness across different domains. We analyze the limitations of existing certified defense strategies and explore the challenges posed by robustness under domain shifts. To address this problem, we introduce a novel causality-inspired framework, GLEAN, designed to learn causal factors that mitigate the negative impact of spurious correlations on robustness, enabling a certifiable defense process across various domains. Extensive experiments on both synthetic and real-world benchmarks verify the effectiveness of our method. GLEAN can pave the path for future work that aims at further exploring causality-inspired defenses and any unified approaches for the generalization of adversarial robustness.

References

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*, 145–155. PMLR.

Anil, C.; Lucas, J.; and Grosse, R. 2019. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, 291–301. PMLR.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, 456–473.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, 387–402. Springer.

Bunel, R. R.; Turkaslan, I.; Torr, P.; Kohli, P.; and Mudigonda, P. K. 2018. A unified view of piecewise linear neural network verification. *Advances in Neural Information Processing Systems*, 31.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehlers, R. 2017. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis: 15th International Symposium, ATVA 2017, Pune, India, October 3–6, 2017, Proceedings 15*, 269–286. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

Jeong, J.; Park, S.; Kim, M.; Lee, H.-C.; Kim, D.-G.; and Shin, J. 2021. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34: 30153–30168.

Jeong, J.; and Shin, J. 2020. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33: 10558–10570.

Katz, G.; Barrett, C.; Dill, D. L.; Julian, K.; and Kochenderfer, M. J. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, 97–117. Springer.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, 5815–5826. PMLR.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, 656–672. IEEE.

Li, B.; Chen, C.; Wang, W.; and Carin, L. 2018. Second-order adversarial attack and certifiable robustness.

Li, B.; Chen, C.; Wang, W.; and Carin, L. 2019. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32.

Li, H.; Wang, X.; Zhang, Z.; and Zhu, W. 2022. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton,

A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Sinha, A.; Namkoong, H.; Volpi, R.; and Duchi, J. 2017. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Sun, J.; Mehra, A.; Kailkhura, B.; Chen, P.-Y.; Hendrycks, D.; Hamm, J.; and Mao, Z. M. 2021. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*.

Sun, J.; Mehra, A.; Kailkhura, B.; Chen, P.-Y.; Hendrycks, D.; Hamm, J.; and Mao, Z. M. 2022. A spectral view of randomized smoothing under common corruptions: Benchmarking and improving certified robustness. In *European Conference on Computer Vision*, 654–671. Springer.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in neural information processing systems*, 33: 1513–1524.

Tjeng, V.; Xiao, K.; and Tedrake, R. 2017. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*.

Trockman, A.; and Kolter, J. Z. 2021. Orthogonalizing convolutional layers with the cayley transform. *arXiv preprint arXiv:2104.07167*.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, 5286–5295. PMLR.

Wong, E.; Schmidt, F.; Metzen, J. H.; and Kolter, J. Z. 2018. Scaling provable adversarial defenses. *Advances in Neural Information Processing Systems*, 31.

Xin, S.; Wang, Y.; Su, J.; and Wang, Y. 2023. On the connection between invariant learning and adversarial training for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10519–10527.

Ye, N.; Li, K.; Bai, H.; Yu, R.; Hong, L.; Zhou, F.; Li, Z.; and Zhu, J. 2022. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7947–7958.

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. *arXiv preprint arXiv:2001.02378*.

Zhang, C.; Zhang, K.; and Li, Y. 2020. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33: 289–301.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *International conference on machine learning*, 819–827. Pmlr.