Generalizability of Adversarial Robustness Under Distribution Shifts

Kumail Alhamoud*, Hasan Abed Al Kader Hammoud*, Motasem Alfarra, Bernard Ghanem
KAUST
{kumail.hamoud, hasanabedalkader.hammoud}@kaust.edu.sa

Abstract

Most existing evaluations of DNN robustness have been done on images sampled from the same distribution on which the model was trained. However, 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. Our study sheds light on the importance of evaluating DNNs under real-world distribution shifts.

1. Introduction

Deep Neural Networks (DNNs) are vulnerable to small and carefully designed perturbations, known as adversarial attacks [21, 9]. 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 [5].

These observations raise concerns regarding the deployability of DNNs in security-critical applications, such as self-driving and medical diagnosis [18, 6, 15]. This brittleness motivated efforts to build models that are not only accurate but also *robust* [10]. Building robust models is usually achieved either **(i)** *empirically*, where the DNN training routine is changed to include such malicious adversarial examples in the training set [16], or **(ii)** *certifiably*, where theoretical guarantees are given about the robustness of a DNN in a region around a given input x [13]. While recent

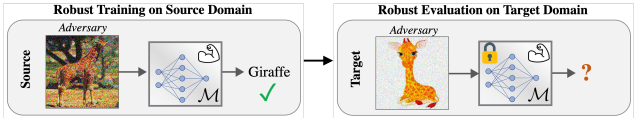

Figure 1. **Does a robust model trained in a (source) domain maintain its robustness when deployed in another (target) domain?** We investigate the generalizability of empirical and certified robustness to various unseen domains.

work on adversarial robustness has made significant strides in developing accurate and robust models, most methods are only evaluated on *in-distribution* data. This means that the training and testing datasets are assumed to be 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 medical image classifier may be trained on data collected from one hospital, but later deployed in a different hospital [2]. Unfortunately, DNNs struggle to generalize to out-of-domain data [7, 8], even in the absence 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 [11]. This setup mimics the unexpected nature of real-world distribution shifts, where models can be regularly exposed to novel domains, and fine-tuning on all these domains becomes impractical. While there has been considerable effort in improving the generalizability of DNNs [22, 20, 17, 25, 19], the generalizability of adversarial robustness to unseen domains remains unexplored.

Our work examines the interplay between domain generalization and adversarial robustness through comprehensive experiments on five standard DG benchmarks provided by DomainBed [11] and WILDS [12]. We investigate 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 [16] on the

*The first two authors contributed equally to this work.

source data. We then inspect the generalizability of certified robustness against input perturbations and parametric deformations by employing Randomized Smoothing (RS) [3] and DeformRS [1]. Our analysis of the generalizability of adversarial robustness to unseen domains 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 necessarily 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 show that empirical and certified robustness generalize to unseen domains in different setups.

2. 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 multi-source domain generalization, there are N source domains of varying sizes $\{D_n\}_{n=1}^{N}$ where for each n, the domain is defined by $D_n = \{(x_j, y_j)\}_{j=1}^{|D_n|} \sim \mathbb{P}_{XY}^{(n)}$. We define the training set S by the union of the N source domains $S = \bigcup_{n=1}^{N} D_n$, and we assume the existence of some unseen target domain $D_{N+1} = \{(x_j, y_j)\}_{j=1}^{|D_{N+1}|} \sim \mathbb{P}_{XY}^{(N+1)}$. 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, which are, in turn, distinct from each other. The aim of DG is to use the source domains S to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the error on the unseen target domain. Since the model is not allowed to sample the target domain during training, 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:

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

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. Section 3 (and 4) investigates the generalizability of empirical (and certified) robustness to diverse target domains.

3. Empirical Robustness and DG

In this section, we study the generalizability of DNNs trained with empirical robustness methods.

Adversarial Training as Augmentation. Adversarial Training (AT) [16] trains the classifier on adversarial examples rather than clean samples. In particular, AT obtains the network parameters θ^* by solving the following optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta, \|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(x+\delta), y) \right], \tag{2}$$

where $\mathcal{D}$ is a data distribution. In general, the inner maximization problem is solved through K steps of Projected Gradient Descent (PGD) [16]. 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 [24] and deploy adversarial training as a data augmentation scheme. In particular, we obtain network parameters θ^* that minimize the following objective:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{P}_{XY}^{(N+1)}} [\lambda \mathcal{L}(f_\theta(x), y) + (1-\lambda)\mathcal{L}(f_\theta(x_{adv}), y)] \tag{3}$$

where $\lambda \in [0, 1]$ controls the robustness-accuracy trade-off. Furthermore, we experiment with the more powerful method TRADES, which has a modified objective [24].

Evaluation Setup. We focus on image classification and adopt the framework of DomainBed [11], which is the standard benchmark in the image domain generalization literature. For each considered dataset, we select a subset of $N-1$ domains to be the source (training) domains and keep the N^{th} domain as the target (evaluation) domain. We follow DomainBed in reporting the average result across all N different source vs. target splits Furthermore, we run each experiment with 3 different seeds and report the standard deviation across our runs. To evaluate the robustness of our models, we assessed their performance on the same norm budget they were trained on. For AutoAttack, we used the default number of steps provided in its implementation. For PGD, we conducted the evaluation with 20 steps.

3.1. Generalization of Empirical Robustness

In Table 1, we report the standard accuracy, ℓ_∞ AutoAttack robust accuracy [4], and ℓ_∞ PGD robust accuracy for nominally- and adversarially-trained models using both TRADES [24] and PGD augmentation Eq. (3) techniques. Four runs with different seeds are averaged, and ϵ is fixed at $2/255$. We address the following questions:

Q1: Does adversarial training improve clean sample generalization in the target domain? Adversarial training methods (TRADES and PGD augmentation) do not enhance clean target accuracy compared to the baseline, as shown in Table 1. ❶ **Unlike transfer learning, where robust training in the source domain is favorable, robust**

Table 1. **Evaluation of ℓ_∞ Robustness.** We assess the robustness of models trained in the source domain by evaluating their clean and robust accuracy in both the source and target domains. Three training approaches are considered: nominal training, PGD adversarial training, and TRADES adversarial training. Robust accuracy is measured against AutoAttack and PGD adversarial attacks. PGD-trained and TRADES-trained models demonstrate transferable robustness to the target distribution.

		Source			Target		
Method	**Dataset**	**Clean Acc.**	**Acc. (AA)**	**Acc.(PGD)**	**Clean Acc.**	**Acc. (AA)**	**Acc. (PGD)**
Baseline	**PACS**	94.69 ± 0.23	1.56 ± 0.33	3.96 ± 0.56	80.24 ± 1.72	0.34 ± 0.16	1.16 ± 0.38
	OfficeHome	77.08 ± 0.30	0.03 ± 0.02	0.40 ± 0.05	58.88 ± 0.85	0.00 ± 0.00	0.56 ± 0.21
	VLCS	84.34 ± 0.12	0.00 ± 0.00	0.00 ± 0.00	74.55 ± 1.04	0.00 ± 0.00	0.02 ± 0.04
	TerraIncognita	86.72 ± 0.22	0.00 ± 0.00	0.00 ± 0.00	44.10 ± 2.52	0.00 ± 0.00	0.00 ± 0.00
PGD	**PACS**	92.83 ± 0.22	76.00 ± 0.33	78.12 ± 0.37	75.29 ± 0.56	56.69 ± 0.91	55.82 ± 1.74
	OfficeHome	72.04 ± 0.40	52.32 ± 0.69	52.81 ± 0.98	52.19 ± 0.73	34.03 ± 0.26	34.51 ± 0.21
	VLCS	79.70 ± 0.22	58.76 ± 0.41	60.02 ± 0.75	69.15 ± 0.74	47.12 ± 0.69	46.73 ± 1.64
	TerraIncognita	71.62 ± 0.74	52.85 ± 2.25	56.05 ± 2.47	27.53 ± 1.45	3.96 ± 1.26	5.59 ± 0.87
TRADES	**PACS**	91.16 ± 0.08	79.89 ± 0.22	79.70 ± 0.95	72.32 ± 0.77	57.96 ± 1.56	57.63 ± 1.45
	OfficeHome	69.12 ± 0.15	54.52 ± 0.74	56.14 ± 0.59	48.47 ± 0.45	35.79 ± 1.14	36.11 ± 1.63
	VLCS	78.58 ± 0.17	63.01 ± 0.79	63.30 ± 0.63	69.36 ± 0.68	52.78 ± 1.66	53.27 ± 0.97
	TerraIncognita	69.81 ± 0.43	58.64 ± 0.79	59.38 ± 0.99	25.27 ± 3.16	5.49 ± 0.59	7.84 ± 0.65

training does not improve clean data accuracy in the target domain without fine-tuning. This contrasts with transfer learning findings and highlights the distinction between transfer learning and domain generalization. Future work should explore the conditions under which adversarial training improves generalization without target fine-tuning.

Q2: Does source domain robustness correspond to target domain robustness? DNNs exhibit reduced robustness when evaluated on distinct target domains. However, ❷ **higher source domain robustness corresponds to higher target domain robustness**. Improving source domain robustness can enhance the out-of-distribution robustness of deployed models, as evidenced by comparing TRADES with PGD in Table 1.

4. Certified Robustness and Domain Generalization

To deploy DNNs in dynamic environments, we need robustness guarantees to carry over into unseen domains. Thus, we study the generalizability of certified robustness.

Randomized Smoothing Background. Randomized smoothing (RS) [3] 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:

$$g_\theta(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \left[f_\theta(x+\epsilon) \right]. \tag{4}$$

Let g_θ predict the 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 [23], g_θ's prediction is certifiably robust at x with certification radius:

$$R = \frac{\sigma}{2} \left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B) \right), \tag{5}$$

where Φ^{-1} is the inverse CDF of the standard Gaussian

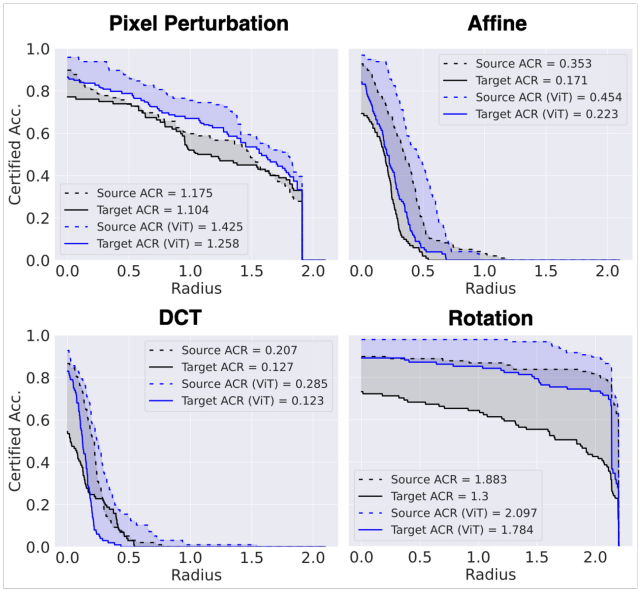

Figure 2. **Generalizability of certified robustness.** We certify ResNet-50 and ViT-Base against pixel perturbations and input deformations. We observe that 1) certified robustness generalizes to unseen domains, and that 2) a stronger architecture (ViT-Base) leads to a better source and target certified accuracy.

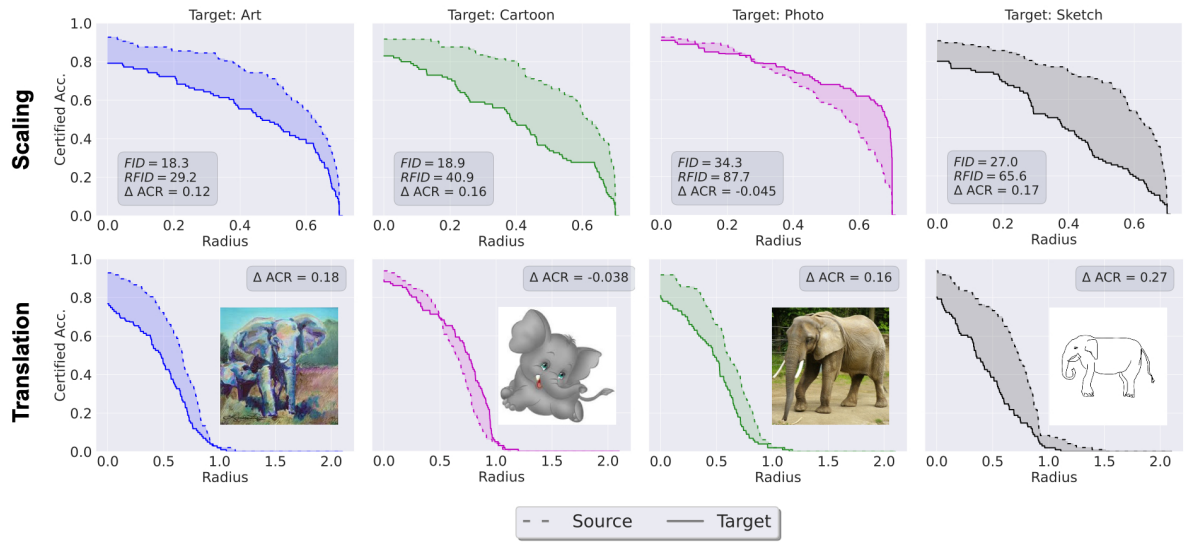

Figure 3. **Does visual similarity correlate with robustness generalizability?** We vary the target domain and plot the certified accuracy curves for two deformations: scaling and translation. A sample from each domain is shown in the second row. The FID/R-FID distances between the source domains and 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.

distribution. As a result of Eq. 5, $\arg\max_i g_\theta^i(x+\delta) = \arg\max_i g_\theta^i(x),\ \forall \|\delta\|_2 \leq R$.

While Eq. 5 provides theoretical guarantees for robustness against additive perturbations, DNNs are also brittle against simple input transformations such as rotation. DeformRS extended randomized smoothing to certify parametric input deformations [1]. In this work, we leverage RS and DeformRS to study the generalizability of certified robustness to unseen target domains.

Experimental Setup. We use the *Photo, Art, Cartoon,* and *Sketch* distributions from PACS [14] to split the data into source and target domains. We employ robust certification techniques, including RS for pixel perturbations and DeformRS for input deformations such as rotation and translation. Data augmentation is applied during training, focusing solely on the source domains. To evaluate certified robustness, we plot accuracy curves for both the source and target domains under various deformations. The certified accuracy at a given radius R represents the percentage of correctly classified test samples with a certified radius of at least R. We calculate the certified radius using established equations and methodologies. Envelope plots are reported to illustrate the best certified accuracy per radius over different smoothing deformation parameters. We estimate p_A and bound p_B using Monte Carlo sampling with 100k samples and a probability of failure of 10^{-3}, following standard practices. Lastly, we compare the ResNet-50 backbone with the ViT-Base transformer model to assess the impact of architecture on generalizability to unseen domains.

4.1. Generalization of Certified Robustness

We investigate the generalizability of certified robustness to unseen domains and explore factors affecting it.

Q3: Does improving the feature extractor enhance target certified accuracy? By switching from ResNet-50 to ViT-Base as the backbone architecture, we observe a significant improvement in target certified robustness across deformations. This aligns with the robustness and domain generalization literature, suggesting that ❸ **stronger backbones lead to better certified robustness and generalization accuracy**.

Q4: Does perceptual similarity correlate with certified robustness generalization? Despite measuring the perceptual similarity between source and target domains using FID and R-FID, we find that ❹ **perceptual similarity metrics are not predictive of performance and robustness generalizability**. Note that *higher* FID/R-FID indicates *less* similarity of distributions. Surprisingly, the photo domain, which has the highest FID and R-FID scores, exhibits the largest certified accuracy generalization.

5. Conclusion

We conducted an extensive empirical analysis of adversarial robustness and domain generalization. We found that empirical and certified robustness generalizes to unseen domains. We also showed that visual similarity is not predictive of the level of generalizability. Based on our findings, we encourage more research on: (i) methods that improve certified accuracy in unseen domains, and (ii) distribution similarity metrics that align with generalization accuracy.

References

[1] M Alfarra, A Bibi, N Khan, P Torr, and B Ghanem. Deformrs: Certifying input deformations with randomized smoothing. In *Proc. of AAAI Conference on Artificial Intelligence*, 2022. 2, 4

[2] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram Van Ginneken, Jeroen Van Der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38, 2019. 1

[3] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 2, 3

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 2

[5] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *ICML*, 2019. 1

[6] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019. 1

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[8] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 1

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 1

[10] Shixiang Shane Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *CoRR*, abs/1412.5068, 2015. 1

[11] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 1, 2

[12] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. 1

[13] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, and Suman Sekhar Jana. Certified robustness to adversarial examples with differential privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2019. 1

[14] Da Li, Yongxin Yang, Yi Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. volume 2017-October, 2017. 4

[15] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021. 1

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2

[17] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[18] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 1

[19] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2021. 1

[20] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016. 1

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014. 1

[22] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance, 2014. 1

[23] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. volume 97, pages 7472–7482. PMLR, 2 2019. 2

[25] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5372–5382, June 2021. 1