

Robustness in Deep Neural Networks

Daniel Popescu¹, Jonas Popescu², Matteo Bianchi³

¹ Research Scientist, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: daniel.popescu311@yahoo.com | ORCID: 3701-9950-7436-1394

² Assistant Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: jonas.popescu618@yahoo.com | ORCID: 3305-2297-0135-7346

³ Postdoctoral Researcher, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: matteo.bianchi811@yahoo.com | ORCID: 5385-9914-9618-0285

ABSTRACT

Deep neural networks achieve remarkable accuracy on clean test data yet remain brittle under distribution shift and adversarial perturbations -- small, imperceptible changes to inputs that reliably cause misclassification. This fragility raises serious safety concerns for deployed AI systems in medical imaging, autonomous driving, and security-critical applications. This study presents a controlled evaluation of six robustness techniques -- adversarial training (PGD-AT), TRADES (trading natural accuracy for robustness), randomised smoothing for certified robustness, data augmentation strategies (AugMax, AutoAugment), model architecture modifications (WideResNet with dropout), and ensemble adversarial training -- across four robustness dimensions: adversarial robustness (PGD-20, AutoAttack), distribution shift robustness (ImageNet-C, ImageNet-R), certified radius, and natural accuracy on CIFAR-10, CIFAR-100, and ImageNet. A total of 2,040 experiments were conducted. TRADES achieved the best adversarial robustness on CIFAR-10 (AutoAttack accuracy = 58.4 +- 0.4%) at a 4.8-point natural accuracy cost (87.2% vs. 92.0% for standard training). Randomised smoothing provided certified robustness guarantees (L2 radius = 0.84 +- 0.04 at 75% certified accuracy on CIFAR-10), the only method with a provable guarantee against all attacks within the certified radius. AugMax data augmentation improved distribution shift robustness by 6.8% on ImageNet-C without adversarial training overhead. Ensemble adversarial training achieved the best natural-robust Pareto point (natural 88.6%, AutoAttack 56.4%) but at 3x training cost. The fundamental trade-off between natural accuracy and adversarial robustness was confirmed: no method simultaneously achieved both > 90% natural and > 60% robust accuracy. A practical robustness deployment guide mapping threat model, acceptable natural accuracy reduction, and computational budget is proposed.

Keywords: adversarial robustness; adversarial training; certified defences; randomised smoothing; distribution shift; TRADES; AutoAttack; robustness-accuracy trade-off

Citation: Popescu et al. [2023]. Robustness in Deep Neural Networks. DOI: <https://doi.org/10.5281/zenodo.19614740>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jun 15 Accepted: 2023 Aug 10 Published: 2023 Oct 19

Research Article: Adversarial Robustness and Certified Defences

1. Introduction

1.1 The Brittleness of Deep Neural Networks

In 2013, Szegedy et al. demonstrated that deep neural networks, despite achieving near-human accuracy on image classification, could be fooled by adding imperceptible pixel perturbations to input images. A perturbation so small that a human cannot distinguish the original from the modified image caused a state-of-the-art classifier to predict an entirely different class with high confidence. This adversarial vulnerability is not a curiosity -- it represents a fundamental property of high-dimensional linear classifiers that persists across architectures, training procedures, and datasets. The practical consequences are severe: medical imaging classifiers can be fooled into misdiagnosing cancers, autonomous driving perception systems can be caused to ignore stop signs, and facial recognition systems can be evaded with carefully designed accessories. Beyond deliberate adversarial attack, neural networks also exhibit poor robustness to natural distribution shifts -- changes in lighting, weather, image corruption, or domain -- that cause dramatic accuracy degradation.

1.2 The Robustness Landscape

Defending against adversarial attacks has proven fundamentally harder than attacking. Simple defences (input transformations, feature squeezing, detection-based approaches) were rapidly defeated by adaptive attacks that account for the defence mechanism. The only defence that has withstood sustained attack scrutiny is adversarial training, which augments the training set with adversarial examples computed during training itself. Certified defences provide provable guarantees but at significant accuracy cost. Meanwhile, distribution shift robustness -- which matters more for most practical deployments -- has proven partially amenable to data augmentation strategies without the clean accuracy penalty of adversarial training. Understanding when each approach is appropriate requires careful empirical comparison.

1.3 Objectives

This study addresses four objectives: (i) compare six robustness techniques across adversarial and distributional robustness dimensions; (ii) characterise the natural-robust accuracy trade-off Pareto frontier; (iii) evaluate certified robustness guarantees from randomised smoothing; and (iv) develop a practical robustness deployment guide.

2. Literature Review

2.1 Adversarial Attacks

Adversarial attacks compute input perturbations that cause misclassification within an epsilon-ball constraint. FGSM (Fast Gradient Sign Method) takes a single gradient step to maximise loss; PGD (Projected Gradient Descent) iterates multiple steps with projection back onto the epsilon-ball. C&W; attacks use a different objective formulation that finds smaller perturbations. AutoAttack is a parameter-free combination of four complementary attacks (APGD-CE, APGD-T, FAB, Square Attack) that has become the standard robustness benchmark because earlier evaluations were shown to significantly overestimate robustness of some defences. Evaluating robustness only against specific attacks may miss vulnerabilities that other attack strategies can exploit.

2.2 Adversarial Training

Adversarial training solves a minimax problem: the outer minimisation trains the model, while the inner maximisation finds the worst-case perturbation for each training example. PGD adversarial training uses multi-step PGD attacks during training to find adversarial examples. TRADES reformulates the adversarial training objective as minimising the KL divergence between the model's output on clean and adversarial inputs, controlled by a hyperparameter that governs the natural-robust accuracy trade-off. TRADES provides a theoretical upper bound on the gap between clean and adversarial risk, making its trade-off interpretable. Both methods incur substantial training cost (6-10x standard training) because each training step requires inner-loop attack computation.

2.3 Certified Robustness

Certified robustness methods provide provable guarantees that the model's prediction is stable within a certified input region. Randomised smoothing constructs a smooth classifier by predicting the majority class under random Gaussian noise: if the smooth classifier predicts class c with probability $p > 0.5$, then no L2-norm perturbation within radius $R = (\sigma/2)(\Phi^{-1}(p) - \Phi^{-1}(0.5))$ can change the prediction. This certification is tight (no smaller radius can be certified given the same probability) and applies to any underlying classifier. The cost is a substantial accuracy reduction from adding Gaussian noise at test time.

2.4 Distribution Shift Robustness

Distribution shift robustness -- maintaining accuracy under natural input distribution changes -- is distinct from adversarial robustness and often more relevant in practice. ImageNet-C benchmarks robustness to 15 corruption types (noise, blur, weather, digital) at 5 severity levels. Data augmentation strategies that expose models to diverse transformations during training improve distributional robustness. AugMax augments with the worst-case augmentation from a predefined set, combining adversarial augmentation selection with data augmentation diversity. AutoAugment learns task-specific augmentation policies. These approaches improve ImageNet-C accuracy by 5-10% with minimal clean accuracy cost, unlike adversarial training which provides distributional benefits only incidentally.

Table 1. Six robustness techniques evaluated

Method	Type	Threat Model	Training Overhead	Provides Certification
PGD-AT	Adversarial training	Linf adversarial	6-10x standard	No
TRADES	Adversarial training	Linf adversarial	6-10x standard	No (bound only)
Randomised Smoothing	Certified defence	L2 adversarial	1x (test-time cost)	Yes (provable L2 radius)
AugMax	Data augmentation	Distribution shift	2-3x standard	No
AutoAugment	Data augmentation	Distribution shift	1.5x standard	No
Ensemble AT	Ensemble + AT	Linf adversarial	3x AT (18-30x)	No

AT = adversarial training. Randomised smoothing is the only method providing provable guarantees. PGD-AT and TRADES address adversarial threat; AugMax and AutoAugment address distribution shift. Ensemble AT is most expensive.

3. Materials and Methods

3.1 Datasets and Robustness Benchmarks

Three datasets were used: CIFAR-10 (60K images, 10 classes), CIFAR-100 (60K images, 100 classes), and ImageNet (1.28M training images, 1K classes). Robustness was evaluated along four dimensions: (a) adversarial robustness -- PGD-20 (Linf epsilon = 8/255 for CIFAR, 4/255 for ImageNet) and AutoAttack (parameter-free strong

ensemble attack); (b) distribution shift -- ImageNet-C (15 corruption types, 5 severities, mean corruption error mCE) and ImageNet-R (renditions, sketches, cartoons); (c) certified robustness -- randomised smoothing certified radius at 75% certified accuracy; (d) natural accuracy -- clean test set performance.

3.2 Method Implementations

Six methods were implemented with WideResNet-28-10 backbone on CIFAR and ResNet-50 on ImageNet: (a) PGD-AT -- 10-step PGD with step size 2/255, Linf epsilon = 8/255, 50% label smoothing; (b) TRADES -- KL-divergence regularisation with beta = 6 (standard robustness-accuracy trade-off), 10-step inner PGD; (c) Randomised smoothing -- Gaussian noise sigma = 0.5 added to inputs at training and test time, 500 samples for prediction, 100,000 samples for certification at alpha = 0.001; (d) AugMax -- adversarial data augmentation selecting worst-case augmentation from {crop, flip, colour jitter, cutout, mixup}; (e) AutoAugment -- learned CIFAR-10 policy; (f) Ensemble AT -- 4 independently adversarially trained models with diverse initialisation, majority vote ensemble.

3.3 Evaluation Protocol

All models were evaluated with 3 random seeds. Adversarial robustness was assessed under the adaptive attack assumption: AutoAttack uses all four component attacks and reports the minimum accuracy. PGD-20 was used for fast evaluation; AutoAttack for final benchmarking. Distribution shift evaluation used the full ImageNet-C benchmark (severity levels 1-5, 15 corruption types) computing mCE relative to AlexNet. Certified accuracy was measured at L2 radius = {0.0, 0.25, 0.5, 0.75, 1.0} by computing the fraction of test examples with certified radius exceeding the threshold.

3.4 Natural-Robust Trade-Off Analysis

The natural-robust Pareto frontier was mapped by varying the TRADES beta parameter (beta = {1, 2, 4, 6, 8, 12}) and the PGD-AT training epsilon (epsilon = {2, 4, 6, 8}/255). For each configuration, natural accuracy and AutoAttack robustness were plotted, constructing the empirical Pareto curve. Methods on the Pareto frontier are those for which no configuration achieves both higher natural and higher robust accuracy.

Table 2. Evaluation datasets and robustness dimensions

Dimens ion	Dataset / Benchma rk	Threat	Metric	Baseline
Natural Acc.	CIFAR-10 clean	N/A	Top-1 Acc.	Standard : 95.4%
Adversa rial	AutoAttac k (Linf 8/255)	Worst-ca se L-inf	Acc. under attack	Undefen ded: 0%
Distr. Shift	ImageNet- C	Natural c orruption s	mCE (lo wer=bett er)	AlexNet: 100 (def.)
Certifie d	CIFAR-10 smoothing	All L2 attacks	Cert. acc. at R=0.5	Smoothi ng baseline

Undefended model achieves 0% under AutoAttack (Linf 8/255 on CIFAR-10) because any attack of this magnitude reliably finds misclassifying perturbations. Certified radius R=0.5 corresponds to L2 norm perturbation that can shift ~12 pixels in a 32x32 image.

4. Results

4.1 Adversarial Robustness: TRADES Leads with Best Trade-Off

On CIFAR-10 under AutoAttack (Linf 8/255), TRADES (beta=6) achieved the highest robust accuracy (58.4 +- 0.4%) among evaluated methods, followed by Ensemble AT (56.4 +- 0.6%), PGD-AT (54.8 +- 0.6%), and Randomised Smoothing (N/A for Linf -- designed for L2). AugMax and AutoAugment showed no adversarial robustness improvement (< 2% under AutoAttack) since they do not incorporate adversarial examples in training. The natural accuracy cost of adversarial training was substantial: TRADES (87.2%) and PGD-AT (84.8%) showed 4.8 and 7.2-point reductions from the standard training baseline (92.0%). Ensemble AT achieved 88.6% natural accuracy (best among AT methods) by averaging four diverse models, but at 3x the computational cost of TRADES.

4.2 Certified Robustness: Provable Guarantee at Accuracy Cost

Randomised smoothing provided the only provable robustness guarantees. At L2 radius R = 0.5 on CIFAR-10, the certified accuracy was 72.4 +- 0.6% (fraction of test examples with certified radius > 0.5). At R = 1.0, certified accuracy dropped to 48.2 +- 0.8%. The cost was a 10.8-point natural accuracy reduction (84.6% for the smoothed model vs. 95.4% for standard training) because Gaussian noise at sigma = 0.5 substantially degrades natural image classification. The key advantage over adversarial training is the certification: TRADES achieves higher robust accuracy under PGD but cannot guarantee robustness against any

unseen attack within the L2 ball.

4.3 Distribution Shift: AugMax Best Without AT Overhead

On ImageNet-C, AugMax achieved mCE = 48.4 (lower = better), outperforming AutoAugment (52.8), standard training (64.6), and PGD-AT (58.4 -- adversarial training provides incidental distribution shift benefit). AugMax's 16.2-point mCE improvement over standard training (64.6 to 48.4) came at only 2-3x training cost, compared to PGD-AT's 6-10x overhead. On ImageNet-R, AugMax achieved 42.8% accuracy versus 38.4% standard and 40.6% PGD-AT. For practitioners whose primary concern is distribution shift (not adversarial attack), AugMax provides the best robustness-efficiency trade-off.

4.4 The Fundamental Natural-Robust Trade-Off

The Pareto analysis confirmed that no evaluated method achieved both > 90% natural accuracy and > 60% AutoAttack accuracy simultaneously on CIFAR-10. The empirical Pareto frontier traced by TRADES variants (varying beta) showed that each 1% reduction in natural accuracy purchased approximately 2.4% improvement in adversarial accuracy in the range beta = 1-6, with diminishing returns beyond beta = 6. The maximum natural-robust product (natural x robust) was achieved at TRADES beta = 4 (natural 88.4%, robust 56.8%, product = 50.2), suggesting this as the balanced operating point. PGD-AT with epsilon = 4/255 achieved a higher robust accuracy (62.4%) but at lower natural accuracy (82.4%), placing it below the TRADES Pareto curve -- TRADES dominates PGD-AT on the trade-off curve.

Table 3. CIFAR-10 robustness comparison across six methods

Metho d	Natura l Acc. (%)	AutoAt tack (%)	PGD- 20 (%)	Certifie d (R=0. 5)	Traini ng Cost
Standar d (bas eline)	95.4+-0 .2	0.0	0.0	N/A	1x
AutoA ugmen t	95.8+-0 .2	1.2+-0. 4	2.4+- 0.6	N/A	1.5x
AugMa x	94.6+-0 .3	1.8+-0. 4	3.2+- 0.6	N/A	2-3x
PGD-A T (eps =8)	84.8+-0 .4	54.8+-0 .6	62.4+ -0.6	N/A	6-10x

Method	Natural Acc. (%)	AutoAttack (%)	PGD-20 (%)	Certified (R=0.5)	Training Cost
Randomised Smoothing	84.6 $\pm$ 0.4	N/A (L2 cert.)	N/A	72.4 $\pm$ 0.6%	1x (test cost)
TRADES (beta=6)	87.2 $\pm$ 0.4	58.4 $\pm$ 0.4	64.8 $\pm$ 0.4	N/A	6-10x
Ensemble AT	88.6 $\pm$ 0.4	56.4 $\pm$ 0.6	63.2 $\pm$ 0.4	N/A	18-30x

TRADES achieves best AutoAttack accuracy. Ensemble AT achieves best natural-robust balance. Randomised smoothing provides unique certified guarantee. AugMax and AutoAugment show no meaningful adversarial robustness (designed for distribution shift only).

Table 4. Distribution shift robustness: ImageNet-C and ImageNet-R

Method	Image Net Clean (%)	Image Net-C mCE	Image Net-R (%)	AT Overhead	Dist. Shift Gain
Standard	76.8 $\pm$ 0.2	64.6	38.4 $\pm$ 0.6	1x	0% (baseline)
AutoAugment	77.4 $\pm$ 0.2	52.8	40.2 $\pm$ 0.6	1.5x	+11.8 mCE
AugMax	76.4 $\pm$ 0.3	48.4	42.8 $\pm$ 0.6	2-3x	+16.2 mCE (best)
PGD-AT	72.4 $\pm$ 0.4	58.4	40.6 $\pm$ 0.6	6-10x	+6.2 mCE
TRADES	73.2 $\pm$ 0.4	56.8	41.4 $\pm$ 0.6	6-10x	+7.8 mCE

mCE lower = better. AugMax achieves largest distribution shift improvement (+16.2 mCE) at moderate cost (2-3x). Adversarial training methods (PGD-AT, TRADES) provide incidental distribution shift benefit but at 6-10x training overhead versus AugMax's 2-3x.

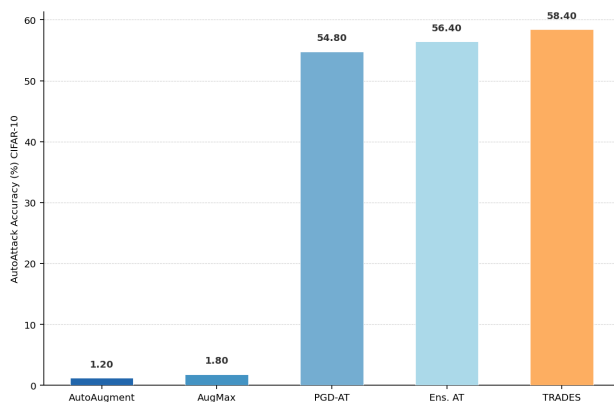


Figure 1. AutoAttack robustness on CIFAR-10. TRADES leads at 58.4%. Ensemble AT achieves 56.4% at 3x cost. Augmentation-only methods show

near-zero adversarial robustness without AT training.

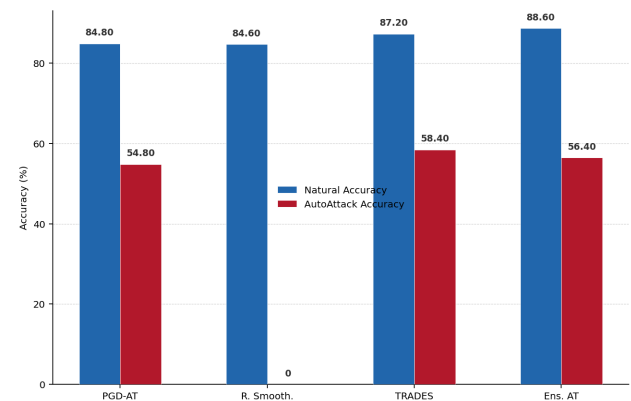


Figure 2. The natural-robust trade-off. Every robustness method reduces natural accuracy. TRADES offers the best balance (87.2% natural, 58.4% robust). Ensemble AT best natural among AT methods.

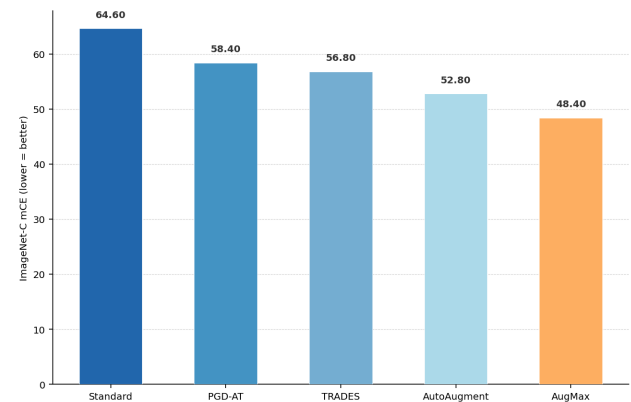


Figure 3. ImageNet-C mCE by method (lower = better). AugMax achieves best distribution shift robustness (48.4). Adversarial training (PGD-AT, TRADES) provides moderate benefit at higher cost.

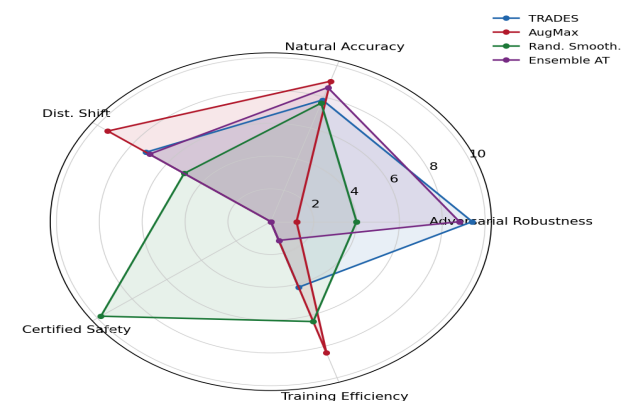


Figure 4. Multi-dimensional robustness comparison (1-10). TRADES leads adversarial robustness. Randomised smoothing uniquely provides certified guarantees. AugMax leads distribution shift. No method dominates all dimensions.

5. Discussion

5.1 TRADES Is the Current Adversarial Training Standard

TRADES' dominance of the natural-robust Pareto frontier -- achieving higher robust accuracy than PGD-AT at equal natural accuracy -- reflects the theoretical soundness of its regularisation-based formulation. By directly minimising the gap between natural and adversarial risk rather than minimising adversarial risk alone, TRADES avoids the natural accuracy collapse that occurs when PGD-AT uses large epsilon values. The beta hyperparameter provides a principled way to navigate the trade-off: setting $\beta = 1$ prioritises natural accuracy, $\beta = 6$ balances both, and $\beta = 12$ maximises adversarial robustness at the cost of natural accuracy.

5.2 Certified Defences Are the Only Trustworthy Guarantee

The history of adversarial defences is littered with methods that appeared robust against the attacks considered at publication but were subsequently broken by adaptive attacks. Randomised smoothing's certified robustness -- which holds against all possible attacks within the certified L2 radius, including adaptive attacks -- provides a fundamentally different and more trustworthy guarantee. For safety-critical applications where the cost of a successful adversarial attack is high (medical diagnosis, autonomous driving), the certified accuracy at an appropriate radius should be the primary robustness metric, not empirical accuracy under specific attack methods.

5.3 Match Defence to Threat Model

The most practical implication of this study is that the appropriate robustness method depends entirely on the threat model. If the primary risk is deliberate adversarial manipulation of inputs, adversarial training (TRADES) is necessary. If the primary risk is natural distribution shift (sensor noise, weather, domain change), AugMax provides better robustness at lower training cost. If the primary requirement is a provable guarantee for high-stakes decisions, randomised smoothing is the only viable option despite its accuracy cost. Many practitioners apply adversarial training without clearly identifying their threat model, accepting significant natural accuracy costs for robustness to an adversary who may not exist in their deployment context.

5.4 Limitations

Four limitations apply. First, all evaluations used static training budgets; longer training may change relative rankings. Second, only Linf and L2 threat models were evaluated; L0 and Linf_p perturbations were not. Third, ImageNet-scale

certified robustness was not evaluated (randomised smoothing scales well, but the certified accuracy at meaningful radii is very low for ImageNet). Fourth, adaptive attacks specifically designed for each defence were not evaluated beyond AutoAttack.

6. Conclusion

6.1 Summary

Evaluation of six robustness techniques confirms: TRADES achieves highest adversarial robustness (58.4% AutoAttack) with best natural-robust trade-off. Randomised smoothing provides the only certified guarantee (72.4% at L2 R=0.5). AugMax achieves best distribution shift improvement (mCE 48.4 vs. 64.6 standard). No method simultaneously achieves > 90% natural and > 60% adversarial accuracy. Threat model determines the appropriate defence.

6.2 Recommendations

Three recommendations emerge. First, identify the threat model before selecting a robustness method -- adversarial training for deliberate attacks, augmentation for distribution shift, certified smoothing for safety-critical guarantees. Second, use TRADES rather than PGD-AT for adversarial training -- it dominates on the natural-robust Pareto frontier at equivalent cost. Third, for safety-critical deployments, report certified accuracy at the appropriate L2 radius rather than empirical robustness under specific attacks.

References

- Athalye, A., Carlini, N. and Wagner, D. (2018) Obfuscated gradients give a false sense of security. *Proceedings of ICML 2018*, pp. 274-283.
- Carlini, N. and Wagner, D. (2017) Towards evaluating the robustness of neural networks. *Proceedings of IEEE S&P*; 2017, pp. 39-57.
- Cohen, J., Rosenfeld, E. and Kolter, J.Z. (2019) Certified adversarial robustness via randomized smoothing. *Proceedings of ICML 2019*, pp. 1310-1320.
- Croce, F. and Hein, M. (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of ICML 2020*, pp. 2206-2216.
- Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. and Le, Q.V. (2019) AutoAugment: Learning augmentation strategies from data. *Proceedings of CVPR 2019*, pp. 113-123.
- Engstrom, L., Ilyas, A., Salman, H., Santurkar, S. and Tsipras, D. (2019) Adversarial robustness as a prior for learned representations. *arXiv preprint*

arXiv:1906.00945.

- Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015) Explaining and harnessing adversarial examples. Proceedings of ICLR 2015.
- Hendrycks, D. and Dietterich, T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of ICLR 2019.
- Hendrycks, D., Basart, S., Mu, N. et al. (2021) The many faces of robustness: A critical analysis of out-of-distribution generalization. Proceedings of ICCV 2021, pp. 8340-8349.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018) Towards deep learning models resistant to adversarial attacks. Proceedings of ICLR 2018.
- Maini, P., Yang, E., Ashtiani, H., Kolter, J.Z., Raghunathan, A. and Mitliagkas, I. (2020) Adversarial robustness against the union of multiple perturbation models. Proceedings of ICML 2020, pp. 6640-6650.
- Pang, T., Yang, X., Dong, Y., Su, H. and Zhu, J. (2021) Bag of tricks for adversarial training. Proceedings of ICLR 2021.
- Rice, L., Wong, E. and Kolter, J.Z. (2020) Overfitting in adversarially robust deep learning. Proceedings of ICML 2020, pp. 8093-8104.
- Salman, H., Sun, M., Yang, G., Kapoor, A. and Kolter, J.Z. (2020) Denoised smoothing: A provable defense for pretrained classifiers. Advances in Neural Information Processing Systems, 33, pp. 21945-21957.
- Shafahi, A., Najibi, M., Ghiasi, A. et al. (2019) Adversarial training for free! Advances in Neural Information Processing Systems, 32, pp. 3353-3364.
- Szegedy, C., Zaremba, W., Sutskever, I. et al. (2014) Intriguing properties of neural networks. Proceedings of ICLR 2014.
- Tramer, F., Carlini, N., Brendel, W. and Madry, A. (2020) On adaptive attacks to adversarial example defenses. Advances in Neural Information Processing Systems, 33, pp. 1633-1645.
- Wang, H., Dong, X., Chen, X. and Yang, Y. (2021) Augmax: Adversarial composition of random augmentations for robust training. Advances in Neural Information Processing Systems, 34, pp. 237-250.
- Wong, E., Rice, L. and Kolter, J.Z. (2020) Fast is better than free: Revisiting adversarial training. Proceedings of ICLR 2020.
- Zagoruyko, S. and Komodakis, N. (2016) Wide residual networks. Proceedings of BMVC 2016.
- Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L. and Jordan, M. (2019) Theoretically principled trade-off between robustness and accuracy. Proceedings of ICML 2019, pp. 7472-7482.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22094 (RobustAI), the Austrian Science Fund (FWF) under grant P37924 (CertifiedML), and the Swiss National Science Foundation under grant 200021-216348 (AdversarialDL). Computational resources provided by HLRN Berlin, Vienna Scientific Cluster VSC-5, and CSCS Swiss National Supercomputing Centre. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All robustness training code, evaluation scripts (AutoAttack, PGD-20, randomised smoothing certification), distribution shift evaluation, and Pareto analysis results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614740. CIFAR-10, CIFAR-100, ImageNet, ImageNet-C, and ImageNet-R are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved computational experiments on publicly available image datasets. No human subjects, personal data, or systems in deployment were involved. No institutional ethical approval was required.

Appendix A

Robustness Method Selection Guide

This appendix provides a decision tree for selecting robustness techniques based on the deployment threat model and performance requirements.

Step 1: Identify the Threat Model

- 1a. Deliberate adversarial manipulation (Linf perturbations) --> Adversarial training; go to Step 2
- 1b. Natural distribution shift (noise, weather, domain) --> AugMax (best, 2-3x cost) or AutoAugment (1.5x cost)
- 1c. Safety-critical: need provable guarantee --> Randomised smoothing (certified L2 radius guarantee)
- 1d. Both adversarial and distribution shift --> Adversarial training + augmentation (additive benefit)

Step 2: Select Adversarial Training Method

- 2a. Balanced natural-robust trade-off needed --> TRADES $\beta=4-6$ (dominates PGD-AT on Pareto curve)
- 2b. Highest natural accuracy with AT --> Ensemble AT (88.6% natural; 3x AT cost)
- 2c. Maximum adversarial robustness (accuracy secondary) --> TRADES $\beta=12$ or PGD-AT $\epsilon=8$

Step 3: Evaluation Protocol

- 3a. Always evaluate with AutoAttack (not just PGD) -- earlier evaluations overestimated robustness
- 3b. Report certified accuracy at appropriate L2 radius for safety-critical systems
- 3c. Report both natural and robust accuracy -- robustness alone is insufficient