

Adversarial Attacks and Certified Defences for Deep Image Classifiers

Rasmi M

Assistant Professor, Department of Computer Applications, Bethania Institute of Management Studies,
Kunnamkulam, India

Article information

Received: 20th May 2026

Received in revised form: 23rd June 2026

Accepted: 25th July 2026

Available online: 12th September 2026

Volume: 1

Issue: 3

DOI: <https://doi.org/10.5281/zenodo.22701174>

Abstract

A change of a few pixel values, too small for a person to notice, can flip the label a deep classifier assigns to an image. Twelve years after that observation, the defensive picture has settled into two camps that are often confused with one another. Empirical defences resist the attacks that were tried against them and carry no guarantee. Certified defences return a proof that no perturbation inside a stated ball can change the prediction, and pay for that proof in accuracy, in compute, or in both. This review covers thirty two works. We describe the attack surface by what the adversary is assumed to know, we trace why a long run of empirical defences was broken by adaptive attacks, and we compare three certification strategies: exact solvers, convex relaxation with bound propagation, and randomized smoothing. Two findings deserve emphasis. Robust accuracy on standard benchmarks remains far below clean accuracy, and the gap is not closing quickly. And an evaluation that does not include an adaptive attack tells the reader almost nothing.

Keywords:- Adversarial Examples, Certified Robustness, Randomized Smoothing, Adversarial Training, Neural Network Verification, Image Classification, Threat Model.

I. INTRODUCTION

Szegedy and colleagues reported something odd in 2014. Take an image a network classifies correctly, add a perturbation chosen by optimisation, and the network changes its answer while a human sees no difference [1]. The perturbation was tiny. The failure was total. That paper started a line of work that is now large enough to need maps rather than summaries.

Goodfellow and colleagues gave the first cheap attack and a first explanation, arguing that the vulnerability comes from local linearity rather than from overfitting [2]. Papernot and colleagues mapped the attack surface [3]. Carlini and Wagner built an optimisation attack strong enough to break defences that had claimed success against weaker ones [4]. Biggio and Roli put the whole thing in the context of a decade of earlier work on adversarial machine learning [14]. Akhtar and Mian catalogued the computer vision case in particular, and the size of their bibliography by 2018 says something about how fast the area grew [15].

The reason to care is not academic. Classifiers now sit inside inspection systems, identity checks and intrusion detection pipelines [31]. An attacker who can move a decision with an imperceptible edit has a capability that ordinary software security does not cover. This review sorts what is known into attacks, empirical defences, and certification, and it states plainly which claims in each category survive scrutiny.

II. THREAT MODELS

Nothing in this field means anything without a threat model. Three things need stating: what the adversary knows, what the adversary can change, and what counts as success. Knowledge runs from full white box access through score based queries to label only queries. Figure 1 lays out the three levels with the attacks that belong to each.

The change is usually bounded by a norm. An L-infinity budget caps how far any single pixel may move, and it dominates the literature because it is easy to optimise against. L-2 budgets cap total energy. Neither corresponds well to human perception, which is a known weakness of the whole formulation. Physical attacks abandon norms altogether: Kurakin and colleagues showed that printed adversarial images survive a camera [13], and Eykholt and colleagues put stickers on road signs and moved a classifier reliably across viewing angles [12].

III. ATTACKS

White box attacks follow the gradient. FGSM takes one signed step [2]. Projected gradient descent iterates that step inside the norm ball and remains the standard baseline [7]. The Carlini and Wagner formulation replaces the constraint with a penalty and solves for the smallest perturbation that changes the label [4]. AutoAttack bundles four diverse attacks, removes the step size choices that let weak evaluations look strong, and has become the default reporting tool [25].

Black box attacks need more work but assume less. The transfer route trains a surrogate model on queried labels, attacks the surrogate, and sends the result to the target [5]. It works often enough to be a practical concern for any deployed image API. Table 1 lists the main attacks with their assumptions.

Table 1. Attacks discussed in this review, with the access each one assumes and the idea it rests on. Descriptions follow the cited papers.

Attack	Access	Core idea	Source
FGSM	White box	One signed gradient step	[2]
PGD	White box	Iterated steps, projected to the ball	[7]
Carlini and Wagner	White box	Minimise perturbation size directly	[4]
AutoAttack	White box	Ensemble with no free parameters	[25]
Surrogate transfer	Black box, labels	Attack a locally trained substitute	[5]
Physical patch	Black box	Printed marks robust to viewpoint	[12], [13]

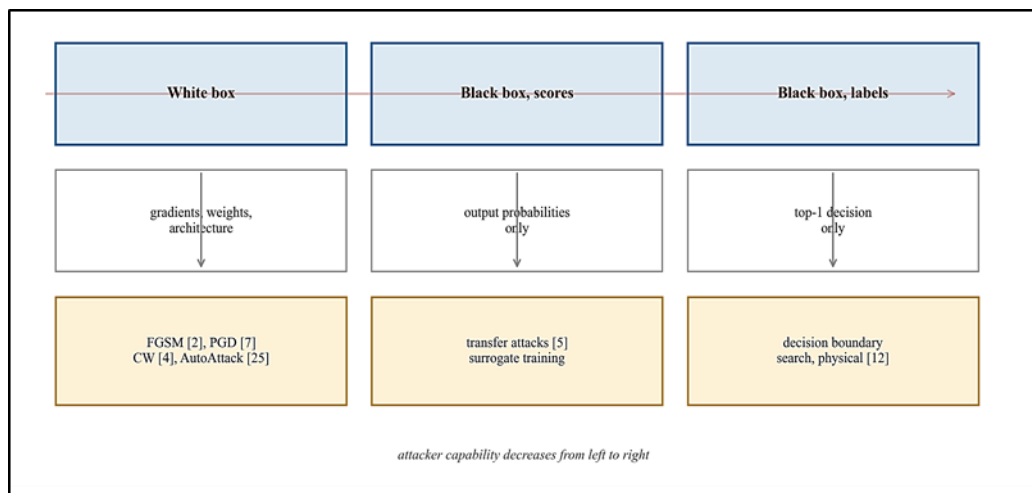


Figure. 1: Threat models ordered by adversary knowledge, with the attacks that operate at each level.

IV. EMPIRICAL DEFENCES, AND WHY SO MANY FAILED

Between 2016 and 2018 dozens of defences were published and then broken, usually within months. Athalye and colleagues explained the pattern in one paper: most of those defences did not remove the adversarial examples, they made the gradient useless to the attacker [8]. Shattered gradients, stochastic gradients and exploding gradients all produce the same symptom, which is that a gradient based attack fails while the vulnerability remains. Replace the gradient with an estimate and the defence collapses.

Tramer and colleagues extended the lesson with a set of case studies, showing that an attack adapted to the specific defence beats a standard attack nearly every time [26]. Detection based schemes such as feature squeezing [11] fall to the same treatment. The practical rule that came out of this period is simple. A defence evaluated only against fixed attacks has not been evaluated.

Adversarial training survived. Madry and colleagues framed training as a saddle point problem and solved the inner maximisation with projected gradient descent [7]. TRADES makes the trade off between clean and robust accuracy explicit in the loss [21]. Neither gives a guarantee, but neither has been broken in the way the gradient masking defences were. Rice and colleagues then found that robust overfitting is severe in this setting, and that early stopping recovers much of what elaborate schemes claim to add [27].

V. CERTIFIED DEFENCES

A. Exact verification

Exact methods answer the question completely. Reluplex extends a simplex solver to handle ReLU constraints [6], and mixed integer formulations do the same through standard optimisation machinery [23]. The answer is sound and complete. The cost grows badly with network size, so this family stays useful for small safety critical networks rather than for image classifiers of realistic depth.

B. Relaxation and bound propagation

Relaxation trades completeness for speed. Wong and Kolter bound the output over a convex outer approximation of the reachable set [9]; Raghunathan and colleagues use a semidefinite relaxation [10]; CROWN propagates linear bounds through general activations [16]; DeepPoly gives an abstract domain for the same purpose [17]. Interval bound propagation is looser still but cheap enough to include inside training, which is how Gowal and colleagues reached certified accuracy on larger images [24]. Beta-CROWN combines bound propagation with branch and bound and closes much of the gap back towards completeness [29].

C. Randomized smoothing

Smoothing takes a different path. Build a new classifier that returns the most likely label under Gaussian noise added to the input, and the resulting function is provably robust in an L-2 ball whose radius follows from the margin between the top two class probabilities [18]. The proof does not depend on the architecture, which is why this family scales to networks that verification cannot touch. Lecuyer and colleagues reached a related result through a differential privacy argument [19], and Salman and colleagues improved the radius by training the base classifier adversarially under noise [20].

The limitations are equally clear. Certification is probabilistic, since the class probabilities are estimated by sampling. Prediction becomes expensive, because each answer needs many noisy forward passes. And the guarantee is L-2 shaped, which does not transfer to the L-infinity threat model that most attacks use. Li and colleagues survey the whole certification landscape and reach similar conclusions [30].

Table 2. Certification strategies compared on the guarantee returned and the reported limits, as described by the cited works.

Strategy	Guarantee	Main limit	Source
Exact solvers	Sound and complete	Network size	[6], [23]
Convex relaxation	Sound, incomplete	Bounds loosen with depth	[9], [10], [16], [17]
Bound propagation in training	Sound, incomplete	Clean accuracy cost	[24], [29]
Randomized smoothing	Probabilistic, L-2 ball	Many forward passes per prediction	[18], [19], [20]

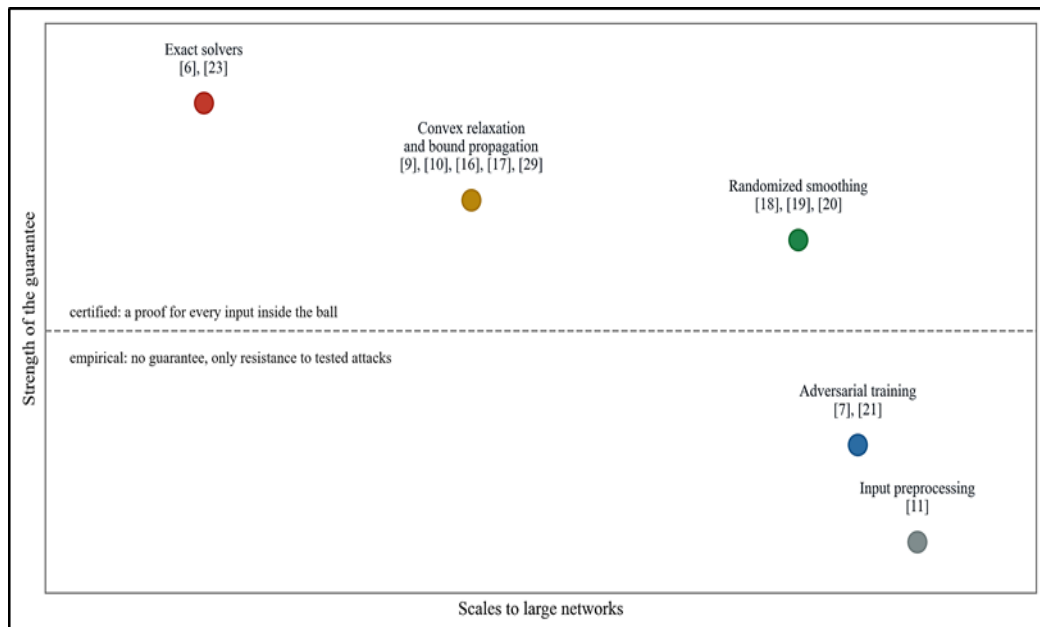


Figure. 2: Defence families placed by guarantee strength and by how far they scale. Positions are qualitative and follow the claims made in the cited papers.

VI. WHAT ROBUSTNESS COSTS

Robustness is not free, and the price shows up in three places. Clean accuracy drops: Tsipras and colleagues argue this is not an artefact of poor training but a property of the problem, since robust and accurate features can genuinely conflict [22]. Training cost rises, because every step of adversarial training runs an inner optimisation. And certified accuracy sits well below empirical robust accuracy, which itself sits well below clean accuracy.

RobustBench collects evaluations under a fixed protocol and makes the size of these gaps hard to ignore [28]. Anyone quoting a robustness number from a single paper should check it there first. Interpretability tools are sometimes proposed as a diagnostic for why a model breaks [32], though explanation and certification answer different questions and should not be substituted for one another.

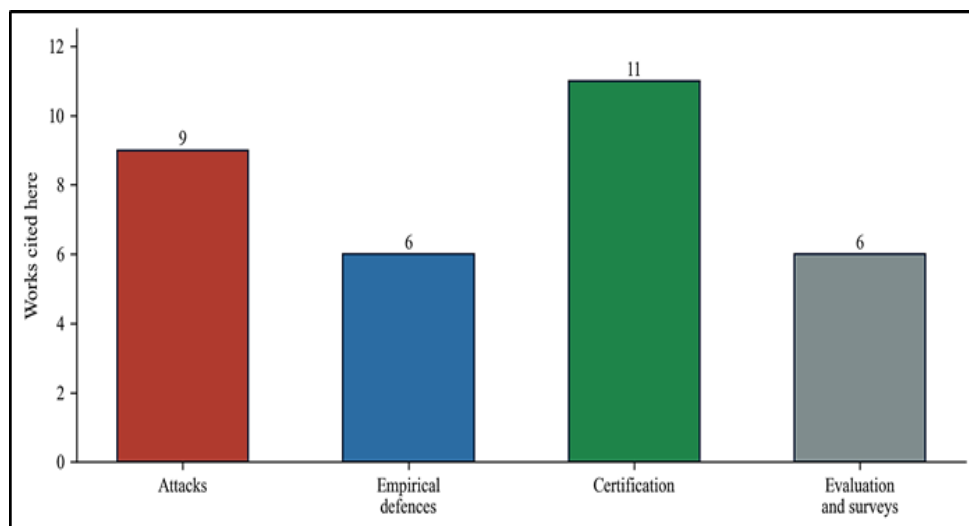


Figure. 3: The thirty two works cited here, coded by the role they play in this review. The chart counts our own reference list and reports no experimental result.

VII. OPEN PROBLEMS

Four gaps stand out. Norm bounded perturbations are a proxy for imperceptibility and a poor one, yet almost the entire literature is built on them. Certification for L-infinity threats at ImageNet scale remains out of reach. Robust overfitting has a diagnosis [27] but no clean cure. And the transfer of all of this to the physical world, where lighting, angle and printing intervene, is studied far less than the digital case [12], [13].

VIII. CONCLUSION

The field has produced one durable empirical method and one durable certified method. Adversarial training with a strong inner attack is the empirical option, and randomized smoothing is the certified option that scales. Everything else either serves smaller networks or has been broken.

For a practitioner the sequence is short. Decide the threat model before writing any code. Evaluate with AutoAttack and at least one attack adapted to your specific defence [25], [26]. Report clean accuracy, empirical robust accuracy and, where a guarantee is needed, certified accuracy with its radius. Papers that skip the third step are describing resistance, not robustness, and the difference matters when a classifier sits in a security path.

REFERENCES

- [1] C. Szegedy *et al.*, “Intriguing properties of neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- [3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in *Proc. IEEE Eur. Symp. Security and Privacy (EuroS&P)*, 2016, pp. 372–387.
- [4] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Security and Privacy (SP)*, 2017, pp. 39–57.
- [5] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proc. ACM Asia Conf. Comput. Commun. Security (AsiaCCS)*, 2017, pp. 506–519.
- [6] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, “Reluplex: An efficient SMT solver for verifying deep neural networks,” in *Proc. Int. Conf. Computer Aided Verification (CAV)*, 2017, pp. 97–117.
- [7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [8] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 274–283.
- [9] E. Wong and J. Z. Kolter, “Provable defenses against adversarial examples via the convex outer adversarial polytope,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5283–5292.
- [10] A. Raghunathan, J. Steinhardt, and P. Liang, “Certified defenses against adversarial examples,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [11] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” in *Proc. Netw. Distrib. Syst. Security Symp. (NDSS)*, 2018.
- [12] K. Eykholt *et al.*, “Robust physical-world attacks on deep learning visual classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1625–1634.
- [13] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Artificial Intelligence Safety and Security*. Boca Raton, FL: Chapman and Hall/CRC, 2018, pp. 99–112.
- [14] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.
- [15] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” *IEEE Access*, vol. 6, pp. 14410–14430, 2018.
- [16] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, “Efficient neural network robustness certification with general activation functions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 4944–4953.
- [17] G. Singh, T. Gehr, M. Puschel, and M. Vechev, “An abstract domain for certifying neural networks,” *Proc. ACM Program. Lang.*, vol. 3, no. POPL, pp. 41:1–41:30, 2019.
- [18] J. Cohen, E. Rosenfeld, and J. Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1310–1320.
- [19] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, “Certified robustness to adversarial examples with differential privacy,” in *Proc. IEEE Symp. Security and Privacy (SP)*, 2019, pp. 656–672.
- [20] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, “Provably robust deep learning via adversarially trained smoothed classifiers,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 11289–11300.
- [21] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 7472–7482.
- [22] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, “Robustness may be at odds with accuracy,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [23] V. Tjeng, K. Y. Xiao, and R. Tedrake, “Evaluating robustness of neural networks with mixed integer programming,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [24] S. Gowal *et al.*, “Scalable verified training for provably robust image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 4841–4850.
- [25] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 2206–2216.
- [26] F. Tramer, N. Carlini, W. Brendel, and A. Madry, “On adaptive attacks to adversarial example defenses,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
- [27] L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 8093–8104.

- [28] F. Croce *et al.*, “RobustBench: A standardized adversarial robustness benchmark,” in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.
- [29] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, and J. Z. Kolter, “Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 29909–29921.
- [30] L. Li, T. Xie, and B. Li, “SoK: Certified robustness for deep neural networks,” in *Proc. IEEE Symp. Security and Privacy (SP)*, 2023, pp. 1289–1310.
- [31] R. N, “AI-driven network security: Using deep learning to detect and mitigate cyber threats in real time,” *Int. J. Inf. Technol. Res. Stud. (IJITRS)*, vol. 1, no. 3, p. 140, Oct. 2025, doi: 10.63090/ijitrs/3139.3209.0012.
- [32] S. E, “Explainable AI (XAI): Enhancing interpretability of deep learning models for critical applications,” *Int. J. Inf. Technol. Res. Stud. (IJITRS)*, vol. 1, no. 2, p. 90, Jul. 2025, doi: 10.63090/ijitrs/3139.3209.0007.