

Adversarial Robustness Evaluation for Deep Learning Models in Safety-Critical Systems

Hariharan P^{1*}, Elangovan S² and Krishnamurthy K³

¹Department of Artificial Intelligence and Data Science, Hemavathi Center for Intelligent Automation, Sri Lanka

Email: poornima.hariharan@hemavathi.edu

ORCID: <https://orcid.org/6877-9737-9899-7720>

²Department of Computer Science and Engineering, Muvattupuzha Center for Computational Cognition, Malaysia

Email: sneha.elangovan@muvattupuzha.edu

ORCID: <https://orcid.org/4795-9930-4552-8819>

³Department of Robotics and Automation, Meenachil Institute of Distributed Intelligence, South Korea Email:

karthik.krishnamurthy@meenachil.edu

ORCID: <https://orcid.org/2678-1008-0180-7318>

ABSTRACT

Deep neural networks now power perception and decision components inside autonomous vehicles, aviation collision-avoidance logic, and clinical decision-support tools, yet the same networks can be driven to confident, silent misclassification by perturbations too small for a human to notice. This article synthesizes the adversarial robustness literature with a specific focus on safety-critical evaluation: the foundational discovery of adversarial examples and the threat model researchers use to reason about them; the attack methodologies, white-box, black-box, and physical-world, that a credible robustness assessment must account for; formal verification and certified-defense techniques that offer provable rather than purely empirical guarantees; and how these ideas have been applied directly to autonomous driving perception, aviation collision-avoidance networks such as ACAS Xu, and clinical imaging models. We give particular weight to the evaluation-methodology literature showing that reported robustness claims frequently collapse under stronger, adaptive attacks, and close with an assessment of where formal verification, physical and multi-modal threat modeling, and certification practice remain unresolved for deployed, large-scale safety-critical systems.

Keywords: adversarial robustness; deep learning safety; adversarial examples; formal verification of neural networks; safety-critical machine learning

Citation: Hariharan P, Elangovan S and Krishnamurthy K [2024]. Adversarial Robustness Evaluation for Deep Learning Models in Safety-Critical Systems. **DOI:** <https://doi.org/10.5281/zenodo.22843887>

Copyright: © 2024 by the authors. Open access under [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Article Information: Received: 8 January 2024 Accepted: 1 April 2024 Published: 9 July 2024

Article Class: Research Article

1. Introduction

Deep neural networks now sit inside systems whose failures are measured in lives rather than in dropped user sessions: perception stacks for autonomous vehicles, collision-avoidance logic for aircraft, diagnostic-support models in radiology and pathology, and increasingly the control loops of industrial and medical robotics. The appeal of deep learning for these roles is obvious — it handles the high-dimensional, unstructured sensor data (camera frames, radar returns, histopathology slides) that classical rule-based systems were never built for, and it does so with accuracy that, on held-out benchmark data, frequently exceeds that of the specialists it is meant to assist. What is less obvious, and considerably more uncomfortable, is that the same networks can be made to fail confidently and silently on inputs that differ from a correctly classified example by a perturbation too small for a human to notice. Szegedy et al. were the first to demonstrate this systematically: gradient-based optimization can find an additive perturbation, imperceptible to a human observer, that flips a network's prediction with high confidence, and — more troubling for anyone hoping this was a quirk of one architecture — that the same perturbation often transfers to other networks trained on the same task [1]. Goodfellow, Shlens, and Szegedy followed with an explanation for why this happens at all: in high-dimensional input spaces, the linear behavior of even nominally nonlinear networks means that small, coordinated perturbations along the gradient direction accumulate into a large change in the network's output, and they showed that a single-step gradient sign method (FGSM) is enough to produce such perturbations cheaply, at scale [2].

That two properties as basic as "small input changes should not flip a confident prediction" and "these changes are easy for an adversary to find" could coexist in a state-of-the-art vision model was, for a research community mostly benchmarking on clean accuracy, a genuine surprise. It has since become one of the more mature subfields of machine learning security, spanning attack algorithms, formal verification techniques, certified defenses, and an increasingly self-critical body of work on evaluation methodology itself — because, as later work would show repeatedly, robustness numbers reported against weak or poorly chosen attacks are not just optimistic, they are frequently wrong by a wide margin [3]. Amodei et al.'s widely cited framing of

concrete problems in AI safety situates adversarial robustness as one instance of a broader pattern the authors call "robustness to distributional shift" — the general problem of a learned system behaving safely when the inputs it encounters at deployment differ from those it was trained and validated on, whether that shift is adversarially engineered or simply an unanticipated operating condition [4]. That framing is a useful one to keep in mind throughout this article: an adversarial perturbation is, in a precise sense, the worst case of a much more general safety problem that any deep learning system deployed outside a tightly controlled validation distribution has to confront. This article surveys the adversarial-specific slice of that broader problem with a specific lens: what does adversarial robustness evaluation actually look like when the deployment target is a safety-critical system, where a misclassification is not an inconvenience but a potential collision, missed diagnosis, or loss-of-separation event?

We organize the discussion in six parts. We first review the foundational discovery of adversarial examples and the threat model researchers use to reason about them. We then survey the attack methodologies — white-box and black-box, digital and physical — that any serious robustness evaluation has to account for, because an evaluation is only as strong as the weakest attack it fails to consider. We turn next to certified and formal verification approaches, which matter disproportionately in safety-critical contexts because they offer guarantees rather than empirical robustness estimates against a finite attack suite. We then look specifically at how these ideas have played out in three safety-critical domains — autonomous driving perception, aviation collision-avoidance, and clinical decision support — where the literature has moved from abstract benchmark demonstrations to attacks and defenses grounded in the actual sensing and decision pipelines these systems use. A section on defenses and evaluation practice follows, including the now well-documented pitfall of gradient obfuscation giving a false sense of security. We close with an honest accounting of open challenges: this is not a solved problem, and treating reported robustness numbers as settled facts, rather than as claims that need to survive a strong adaptive attack, remains one of the more consequential mistakes a safety-critical systems engineer can make.

It is worth being explicit up front about the scope of this synthesis. This is a review of published, peer-reviewed and widely cited work, not a report of new experiments; every specific number cited below is drawn from the source paper reporting it, and we have verified each one against the original publication rather than repeating it from secondary summaries. Where the literature itself is candid about a limitation — and it frequently is — we say so rather than smoothing it over.

2. Foundations: The Discovery and Threat Model of Adversarial Examples

Szegedy et al.'s original demonstration used L-BFGS to find the smallest perturbation, measured in an L2 sense, that would cause a trained network to misclassify an image while remaining visually indistinguishable from the original to a human [1]. Two findings from that paper have shaped essentially all subsequent work. First, the perturbations required were surprisingly small — often imperceptible under normal viewing conditions — which ruled out the comforting hypothesis that adversarial examples were simply out-of-distribution inputs that happened to look unusual. Second, and more consequential for threat modeling, the perturbations transferred: an adversarial example crafted against one network frequently fooled a second network trained independently on the same task, even with a different architecture. Transferability means an attacker does not need white-box access to the deployed model to mount an attack; a substitute model trained on similar data, or even just knowledge of the task, can be enough, which is precisely the black-box threat model that later work formalized more rigorously [5].

Goodfellow, Shlens, and Szegedy's Fast Gradient Sign Method (FGSM) reframed the phenomenon in a way that made it both cheaper to exploit and easier to reason about theoretically. Rather than an expensive iterative optimization, FGSM takes a single step in the direction of the sign of the gradient of the loss with respect to the input, scaled by a small constant [2]. The authors argued that adversarial vulnerability is not a symptom of over-fitting or excessive nonlinearity, as had been assumed, but a natural consequence of models that are locally approximately linear in high-dimensional input space: a perturbation that nudges every pixel a tiny amount in a consistent, gradient-aligned direction can produce a large change in the pre-softmax logits even though no individual pixel change is perceptible. This

linear-behavior explanation also predicted, correctly, that adversarial training — explicitly including FGSM-perturbed examples in the training set — could measurably improve robustness against that same attack, a defense strategy that both this paper and much later work built on.

Kurakin, Goodfellow, and Bengio extended the threat model from the digital domain, where an attacker can write directly to a model's input tensor, to the physical domain, where an attacker can only manipulate what a camera later photographs [6]. They showed that adversarial examples generated digitally, and then printed and re-photographed with a phone camera, retained much of their adversarial effect despite the additional noise introduced by printing, lighting, and camera sensor variation, and they proposed an iterative variant of FGSM to make attacks more robust to this pipeline. This is a foundational result for safety-critical robustness evaluation specifically, because it establishes that the threat model for a vision system reading physical signals — a road sign, a printed label, a patient scan displayed on a monitor — cannot be dismissed as a purely digital, laboratory-only concern; the perturbation only has to survive the physical acquisition process, not be applied directly to the model's input tensor.

Taken together, this foundational body of work established the basic threat model that essentially every later robustness evaluation inherits: an adversary, with some level of access (white-box gradient access, black-box query access, or no access beyond a substitute model), seeks a perturbation within some bounded norm ball (L, L2, or a physically realizable constraint set) that causes a target model to misclassify, while the perturbation itself remains within whatever constraint — imperceptibility, physical realizability, or a fixed budget — defines the attack's threat model. Robustness evaluation is, at its core, the empirical or formal question of how large that perturbation must be, under a given norm and threat model, before an attack succeeds.

3. Attack Methodologies for Robustness Evaluation

A robustness claim is only as strong as the attacks used to test it, and the attack literature has grown considerably more sophisticated than a single FGSM pass. Moosavi-Dezfooli, Fawzi, and Frossard's DeepFool iteratively linearizes the decision boundary around an input and steps toward the nearest boundary in each

iteration, producing near-minimal L2 perturbations far smaller than FGSM typically finds, which makes it a useful tool for measuring a model's actual margin to misclassification rather than just confirming that *some* adversarial example exists [7]. Madry et al.'s projected gradient descent (PGD) attack, developed as part of a broader robust-optimization framing of adversarial training, iterates FGSM-style steps while projecting back onto the allowed perturbation ball after each step, and the authors argued — with empirical support that has held up remarkably well in subsequent evaluation practice — that PGD with enough random restarts and iterations is close to the strongest first-order (gradient-based) attack achievable within a given L budget, making it close to a universal adversary for evaluation purposes [8].

Carlini and Wagner's optimization-based attacks (commonly abbreviated C&W) were developed explicitly as a response to a wave of defenses that had reported strong robustness against FGSM and other weaker attacks; C&W formulates the attack as a constrained optimization problem solved with a change-of-variables trick that avoids the box-constraint difficulties of naive gradient descent on pixel values, and the paper showed that several defenses claiming near-complete robustness collapsed almost entirely once evaluated against this stronger attack [9]. This paper is arguably as important methodologically as it is technically: it established the now-standard practice of treating a defense's headline robustness number with suspicion until it has been tested against the strongest attack the evaluator can construct with knowledge of the defense mechanism, rather than against whatever attack happened to be reported in the defense's own paper.

Papernot et al.'s earlier work on the limitations of deep learning in adversarial settings introduced a complementary white-box attack built around the forward derivative (the Jacobian of the network's output with respect to its input) to construct adversarial saliency maps identifying which input features have the largest influence on a target misclassification, allowing an attacker to perturb only a small, targeted subset of features rather than the whole input — a useful contrast to FGSM's whole-image, single-step perturbation and a reminder that "adversarial perturbation" covers a family of quite different attack geometries, not one fixed recipe [10]. Not every attacker has white-box gradient access to

the deployed model, however, and Papernot et al.'s later practical black-box attack addressed exactly this constrained setting: an attacker trains a local substitute model using synthetic queries to the target model's prediction API, uses Jacobian-based dataset augmentation to efficiently approximate the target's decision boundary near the training points, and then crafts adversarial examples against the substitute, relying on the transferability property Szegedy et al. first observed to have them succeed against the black-box target [5]. The paper demonstrated this attack succeeding against real, deployed machine-learning-as-a-service APIs, which matters directly for safety-critical systems procurement: a vendor's claim that a model is "protected" because its weights are not public does not, by itself, rule out a successful attack, since query access alone is often sufficient.

Physical-world attacks push the threat model further still, and Eykholt et al.'s work on robust physical perturbations is the reference point for this line of research in the context most relevant to autonomous driving: they showed that carefully designed stickers, placed on a stop sign following an optimization procedure explicitly designed to survive varying viewing distances, angles, and lighting conditions, caused a standard road-sign classifier to misread the sign as a speed-limit sign with high reliability across a range of physical test conditions, not merely in simulation [11]. Brown et al.'s Adversarial Patch takes the physical-attack idea to its logical extreme: rather than a small, imperceptible perturbation, a localized, visually obvious patch is optimized to dominate a classifier's decision regardless of the rest of the scene content, and the resulting patches were shown to transfer across models, scales, and placements when printed and photographed [12]. For safety-critical perception, the patch-attack literature is a reminder that "small perturbation" is not the only threat model worth evaluating against; a localized but visually salient attack that a human would notice but a downstream automated system would not flag as anomalous is arguably a more realistic threat in many physical deployment contexts than an imperceptible whole-image perturbation.

Cao et al.'s adversarial sensor attack extends the physical threat model beyond cameras to LiDAR, the depth-sensing modality many autonomous vehicle

stacks rely on for obstacle detection; the authors demonstrated a spoofing attack, using a custom-built device to inject carefully timed laser pulses, that could fool an actual production-grade 3D object detection pipeline into either hiding a genuine obstacle or hallucinating a nonexistent one, evaluated against real vehicle-grade LiDAR hardware and a widely used detection model [13]. This paper is a useful corrective to a vision-camera-centric reading of the adversarial robustness literature: sensor fusion and multi-modal perception, often proposed as a mitigation for camera-only attacks, are themselves subject to modality-specific adversarial threats that a camera-focused robustness evaluation would simply miss.

4. Certified and Formal Verification of Robustness

Empirical attack-and-defend evaluation, however thorough, has an inherent limitation for safety-critical certification: it can demonstrate that a specific attack suite fails to break a model, but it cannot prove that no attack within a given threat model exists. This gap motivates a distinct line of research aimed at formal guarantees. Katz et al.'s Reluplex is the foundational result here: an SMT (satisfiability modulo theories) solver extended with a specialized decision procedure for the piecewise-linear ReLU activation function, capable of proving properties of neural networks — including local adversarial robustness, that no input within a bounded region around a given point changes the network's output classification — for networks with hundreds of neurons, which was a substantial scaling advance over prior verification approaches at the time [14]. Critically, the paper's central case study was not an abstract benchmark network but the ACAS Xu airborne collision-avoidance system, a safety-critical aviation control policy compressed into a neural network, and Reluplex was used to formally verify properties such as "if an intruder aircraft is near and approaching from the left, the system will not advise a turn to the right" — precisely the kind of correctness property that certification authorities require for airborne systems and that empirical testing alone cannot establish with certainty.

Huang, Kwiatkowska, Wang, and Wu approached the same underlying problem — establishing formal guarantees on neural network behavior under input

perturbation — from an SMT- and search-based verification angle that emphasizes exhaustively covering the space of perturbations up to a given resolution and norm bound, and framing safety verification as a certification problem: given a network and a region of input space, either establish that the network's classification is invariant throughout that region, or exhibit a counterexample [15]. Their approach and Reluplex's differ in emphasis and scalability tradeoffs, but both established, at roughly the same moment in the literature, that adversarial robustness verification for neural networks was a tractable formal-methods problem for networks of meaningful (if still modest, by today's standards) size, opening a research direction that has since scaled considerably further.

Table 1: Representative formal-verification and certified-robustness approaches relevant to safety-critical evaluation, with the guarantee type and reported result each cited source gives. [14,15,16,17,18].

Approach	Source	Guarantee Type	Reported Result
Reluplex (SMT solver with ReLU decision procedure)	Katz et al. [14]	Exact formal verification	Formally verified ACAS Xu collision-avoidance safety properties on networks with hundreds of neurons
SMT/search-based safety verification	Huang et al. [15]	Exact formal verification	Proves classification invariance or produces a counterexample within a bounded input region
Convex outer adversarial polytope	Wong and Kolter [16]	Certified bound via convex relaxation	Trains networks with a provable upper bound on worst-case adversarial loss within an L-infinity ball
Randomized smoothing	Cohen et al. [17]	Probabilistic certified radius	49% certified top-1 accuracy on ImageNet under L2 perturbations up to radius 0.5
AutoAttack (parameter-free attack ensemble)	Croce and Hein [18]	Standardized empirical lower bound	Ensemble of 4 diverse attacks used to reveal overstated robustness across many published defenses

It is worth pausing on why exact, SMT-based verification and certified, relaxation-based defenses are treated as a genuinely distinct category from the attack-and-defend empirical literature surveyed in the previous section, rather than simply a stronger flavor of the same activity. An empirical attack, however strong, only ever demonstrates that a model fails on the specific inputs the attack happened to find; it cannot, on its own,

demonstrate that no adversarial input exists within a given perturbation budget, because the search space of possible perturbations is generally far too large to exhaust by direct trial. A formal verification result, by contrast, is a statement about every point in a bounded region of input space simultaneously, established through the network's own mathematical structure rather than through search. This distinction maps directly onto the two different questions a safety-critical certification process actually needs answered: "has this system been shown to fail under the attacks we tried" (an empirical, falsifiable claim that attack-and-defend evaluation answers) versus "can this system be shown, exhaustively, never to fail within a stated envelope of adversarial input" (a formal claim that only verification, not attack search, can in principle establish). Reluplex and the safety-verification work that followed it were significant specifically because they showed the second, stronger kind of question was tractable at all for realistic (if still modest) neural network sizes, rather than being a purely theoretical possibility with no practical instantiation [14,15].

Formal, exact verification remains computationally expensive as network size grows, which motivated a parallel line of work on certified defenses that trade exactness for scalability. Wong and Kolter's convex outer adversarial polytope approach bounds a network's worst-case behavior under bounded input perturbation using a convex relaxation of the network's nonlinearities, computed via a dual optimization problem, and trains the network to explicitly minimize this provable upper bound on adversarial loss rather than only its empirical loss on attacked examples; the resulting networks come with a certificate — a provable guarantee, not just an empirical observation — that no attack within the specified L_2 ball can change the classification for a given input [16]. Cohen, Rosenfeld, and Kolter took a different, sampling-based route to certification with randomized smoothing: given any base classifier, construct a new "smoothed" classifier that returns the class most likely to be predicted when Gaussian noise is added to the input, and derive a certified L_2 radius, using a tight, information-theoretic argument, within which the smoothed classifier's prediction is provably constant [17]. The appeal of randomized smoothing for safety-critical deployment is that it is architecture-agnostic — it wraps essentially any classifier, including large-scale ImageNet models,

without requiring the specialized relaxation-based training convex-polytope methods need — at the cost of a certified radius that is generally smaller and a probabilistic (rather than exact) guarantee, since the certification itself relies on a statistical argument with a small failure probability.

Because certified defenses generally trade off robustness radius, accuracy, and computational cost against each other, and because a certified defense that is never attacked in practice tells you little about how it behaves against attacks outside its certified guarantee, empirical attack evaluation and formal certification are best read as complementary rather than substitutable. This is where Croce and Hein's AutoAttack becomes directly relevant to certified-defense evaluation as well as to empirical defense evaluation: AutoAttack is an ensemble of four diverse, parameter-free attacks — including two enhanced PGD variants with different loss functions and step-size schedules, a query-efficient black-box attack, and a decision-based boundary attack — assembled specifically to remove the human-tuning variable from robustness evaluation, since the authors documented case after case in the published literature where a defense's reported robustness collapsed once evaluated with better-tuned attack hyperparameters than the defense's own paper had used [18]. AutoAttack has since become close to a de facto standard benchmark (via the associated RobustBench leaderboard) for reporting empirical robustness in a way that is comparable across papers, precisely because it removes much of the researcher-degrees-of-freedom problem that had made earlier reported robustness numbers difficult to compare or trust.

5. Adversarial Robustness in Safety-Critical Domains

The general attack and verification literature surveyed above has, over the past several years, been applied with increasing specificity to three domains where the stakes of a misclassification are unambiguous: autonomous driving perception, aviation collision-avoidance, and clinical decision support.

Autonomous driving perception has received the most sustained attention, in part because its camera- and LiDAR-based perception stacks are directly descended from the same computer-vision architectures the foundational adversarial examples literature targeted. Eykholt et al.'s physical stop-sign attack and Cao et al.'s

LiDAR spoofing attack, both discussed above, demonstrate that the threat is not confined to simulation: both papers evaluated against physical hardware and, in Cao et al.'s case, against a production-representative detection pipeline, producing misclassifications and missed or hallucinated detections that would directly translate into unsafe driving decisions if left unmitigated [11,13]. The practical implication for a safety-critical evaluation process is that camera-only, digital-perturbation-only threat models are insufficient; a credible adversarial robustness assessment for an autonomous vehicle perception stack needs to account for physically realizable perturbations across every sensing modality the stack fuses, evaluated against the actual downstream detection and tracking pipeline rather than an isolated classifier.

Aviation offers a smaller but methodologically distinctive body of work, largely because the domain's certification culture (DO-178C and related standards for airborne software) puts a premium on formal guarantees over empirical robustness claims. Julian et al.'s work compressing the ACAS Xu collision-avoidance lookup tables into a neural network policy — replacing gigabytes of tabular data with a compact network while preserving decision quality — created exactly the artifact that motivated Katz et al.'s Reluplex verification work: a safety-critical control policy small enough to formally verify but expressive enough to be genuinely useful in place of the original lookup tables [19,14]. The ACAS Xu case study is, in our reading, the clearest existing example in the literature of the full pipeline a safety-critical adversarial-robustness evaluation should aspire to: a neural network deployed in place of a previously verified traditional system, subjected to formal verification against explicitly stated safety properties, with counterexamples (where found) used to iterate on the network before deployment, rather than robustness being assessed only empirically after the fact.

Clinical machine learning presents a different flavor of the same underlying problem, because the "attacker" in Finlayson et al.'s treatment is not necessarily a malicious external party but potentially includes actors with an economic incentive internal to the healthcare system itself — providers or institutions seeking to alter a model's output to influence billing, insurance authorization, or reported outcomes [20]. The authors

demonstrated, across dermatology, ophthalmology, and radiology models, that small, often imperceptible perturbations to medical images could change a model's diagnostic classification, and argued that the healthcare system's specific incentive structures (billing codes tied to diagnoses, for instance) make adversarial manipulation of clinical models a more immediately plausible threat than in many other domains, not a hypothetical one requiring an unusually sophisticated external adversary. This is a useful corrective to a narrow reading of "safety-critical" as solely about accidental, environmentally caused failures; deliberate, economically motivated adversarial manipulation is itself a safety concern when the manipulated output determines a patient's diagnosis or treatment.

6. Defenses, Evaluation Pitfalls, and the Obfuscated-Gradients Problem

Adversarial training — including adversarially perturbed examples during training so the model learns to classify them correctly — remains the most empirically durable defense in the literature, and Madry et al.'s PGD-based adversarial training, framing the training objective as a min-max robust optimization problem (minimize loss under the worst-case perturbation an inner-loop PGD attack can find), has held up against a wide range of subsequent attacks better than most defenses proposed before or since [8]. Zhang et al.'s TRADES refined this idea by explicitly decomposing the robust loss into a natural (clean-accuracy) term and a boundary term that penalizes disagreement between the model's predictions on clean and adversarially perturbed versions of the same input, giving practitioners an explicit, tunable knob on the accuracy-robustness tradeoff that plain PGD adversarial training does not directly expose, and reporting a first-place result in the NeurIPS 2018 Adversarial Vision Challenge as empirical validation of the approach [21]. Tramèr et al.'s ensemble adversarial training addressed a specific weakness of single-model adversarial training: a model adversarially trained against its own gradient can still be vulnerable to a degenerate solution where gradients near the data manifold become uninformative rather than genuinely robust, and the paper showed that augmenting training data with adversarial examples transferred in from a diverse set of independently trained (and frozen, non-attacked) models produces a more genuinely robust decision boundary, a defense that won the first round of the NeurIPS 2017 Competition on Defenses

Against Adversarial Attacks [22].

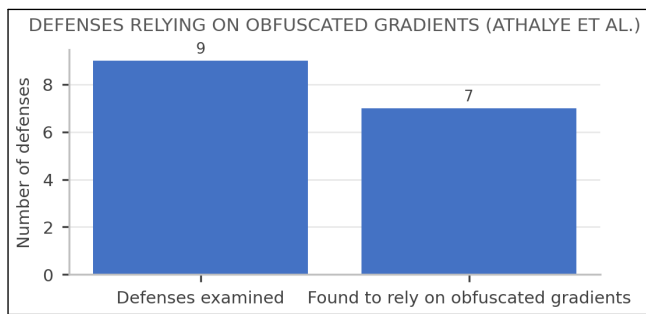


Figure 1: Of nine non-certified white-box defenses accepted at ICLR 2018, Athalye et al. found that seven relied on some form of gradient obfuscation rather than genuine robustness, and reduced their reported robust accuracy to near zero once attacked with methods designed to see through the obfuscation -- the finding that motivated adaptive, ensemble-based evaluation standards such as AutoAttack [3].

Xie et al.'s feature denoising approach targets a different point in the pipeline: rather than modifying the training loss, the architecture itself is augmented with denoising blocks (non-local means or related filtering operations) inserted between convolutional layers, explicitly designed to suppress the noise-like patterns adversarial perturbations introduce into intermediate feature maps, combined with adversarial training at ImageNet scale; the authors reported that this combination substantially improved robustness against strong white-box attacks on ImageNet-scale models relative to adversarial training alone, at a scale (ImageNet, rather than CIFAR-scale benchmarks) that much of the earlier defense literature had not been tested at [23]. Scale matters here specifically for safety-critical relevance, since production perception systems in automotive and medical settings typically operate on higher-resolution, higher-complexity inputs than the small benchmark datasets (CIFAR-10, MNIST) that dominate the academic defense literature.

No discussion of defenses is complete without confronting Athalye, Carlini, and Wagner's obfuscated-gradients paper directly, because its central finding reshaped how the entire field evaluates robustness claims. The authors examined a set of defenses accepted at a major venue (ICLR 2018), each reporting strong robustness against standard attacks, and showed that the overwhelming majority relied, whether the authors intended it or not, on some form of gradient obfuscation — shattered gradients, stochastic gradients, or exploding/vanishing gradients that make standard gradient-based attacks fail not because the model is genuinely robust but because the attack's

optimization signal is degraded or misleading [3]. Using attack techniques specifically designed to see through each obfuscation mechanism (backward pass differentiable approximation, expectation over transformation, and reparameterization, among others), they reduced the reported robust accuracy of seven of nine examined defenses to near zero. This paper's methodological legacy is arguably larger than any single attack or defense technique surveyed above: it established that a defense's robustness claim cannot be taken at face value from a single reported attack-accuracy number, and that a defense's authors — no matter how carefully they evaluated against existing attacks — cannot be assumed to have anticipated every adaptive attack a determined adversary with knowledge of the defense mechanism would construct. Croce and Hein's AutoAttack, discussed above, is in significant part a direct institutional response to this problem: a standardized, diverse attack ensemble specifically designed to catch the failure modes obfuscated-gradients defenses exhibit, so that future robustness claims are harder to overstate by accident [18].

Two further pieces of work reframe rather than directly attack-or-defend the problem, and both are relevant to how a safety-critical systems engineer should interpret the whole literature above. Ilyas et al. argued that adversarial examples are not simply a bug or an artifact of insufficiently regularized training, but arise from models genuinely learning to exploit brittle, human-imperceptible but statistically predictive features present in the training distribution — features that are "real" in the sense of being predictive of the label on the clean data distribution, but not robust to small perturbations, and not aligned with the features a human uses to make the same classification [24]. This reframing has a direct practical implication for safety-critical deployment: robustness and accuracy are not simply two independently improvable metrics, because a model trained purely to maximize clean accuracy is, on this account, actively incentivized to rely on non-robust features, which means robustness has to be an explicit training and evaluation objective, not an emergent byproduct of sufficient accuracy or scale. Biggio and Roli's retrospective, "Wild Patterns," surveys the decade of adversarial machine learning research (predating even the deep-learning-specific discovery by Szegedy et al., with roots in adversarial attacks on linear

classifiers for spam and malware detection) and makes a related, historically grounded argument: that treating adversarial robustness as a deep-learning-specific problem, rather than a general property of any learning system deployed in an adversarial or safety-critical environment, causes each new subfield to rediscover the same lessons about the gap between benchmark performance and adversarial reality [25]. Akhtar and Mian's contemporaneous survey and Xu et al.'s later review both catalogue the breadth of this literature across attack types, defense families, and application domains (including, in Xu et al.'s case, graph neural networks and text models beyond the vision-centric focus of most of the papers discussed above), and are useful entry points for a reader who wants the fuller taxonomy that this article's safety-critical focus necessarily leaves out [26,27].

7. Open Challenges

Several gaps recur across this literature and deserve to be stated plainly rather than smoothed over. First, the scale mismatch between formal verification and production models remains substantial: Reluplex-style exact verification, and even the relaxation-based certified defenses that followed it, remain computationally demanding as network size and input dimensionality grow, and the ACAS Xu networks Katz et al. verified are orders of magnitude smaller than the convolutional and transformer-based perception models deployed in production automotive and clinical systems today [14,16]. A robustness certificate for a compact aviation policy network does not straightforwardly extend to a guarantee for a large-scale vision transformer processing raw camera frames, and the field has not fully closed this gap.

Second, physical-world and multi-modal threat models remain comparatively under-explored relative to the digital, single-modality attack literature. Eykholt et al.'s stop-sign work and Cao et al.'s LiDAR spoofing work are notable precisely because they are exceptions to a broader pattern in which most reported robustness evaluations still default to L or L2-bounded digital perturbations on a single input modality, evaluated in isolation from the sensor fusion, tracking, and planning logic that a real autonomous system would run downstream of a compromised perception module [11,13]. A perception model that is robust in isolation may still contribute to an unsafe system-level outcome if

the fusion and planning logic downstream trusts its output uncritically, and system-level (rather than model-level) robustness evaluation remains comparatively immature in the published literature. This point is easy to state and hard to operationalize, because it requires evaluating robustness against the full sensing-to-actuation pipeline rather than a single model checkpoint, which in turn requires the kind of end-to-end, hardware-in-the-loop testing infrastructure that most academic robustness research, understandably, does not have access to.

Third, the evaluation-methodology problem Athalye et al. identified has not disappeared simply because it is now well known; it has become somewhat more subtle [3]. AutoAttack and RobustBench have substantially raised the floor for what counts as a credible robustness claim in the mainstream image-classification defense literature, but domain-specific defenses in safety-critical applications — a bespoke defense for a LiDAR pipeline, or a clinical-imaging-specific preprocessing defense — do not always have an equivalent standardized, adaptive attack suite available off the shelf, and a defense evaluated only against generic attacks borrowed from the image-classification literature risks repeating exactly the mistake the obfuscated-gradients paper diagnosed, just in a domain where the standard remedy has not yet been fully worked out.

Fourth, the tension Ilyas et al. identify between clean accuracy and robust reliance on non-robust-but-predictive features has direct regulatory and certification implications that the technical literature has only begun to grapple with [24]. Safety-critical system certification processes (in aviation, medical devices, and increasingly automotive functional-safety standards) are built around demonstrating reliable, explainable behavior, and a model whose clean-data accuracy depends partly on features that are, by construction, not robust to small perturbations sits uneasily within that certification paradigm — not because the model is necessarily unsafe in practice, but because the standard tools for demonstrating safety were not designed with this specific failure mode in mind.

None of these gaps argue against using deep learning in safety-critical systems altogether; the literature synthesized here, taken as a whole, argues for something more specific — that adversarial robustness evaluation

needs to be integrated into the safety-case and certification process as a first-class concern, using the strongest available attacks (not the attacks a defense's own authors happened to test against), formal verification where computationally feasible, and an explicit accounting of the physical, multi-modal, and system-level threat surface a deployed model will actually face, rather than the narrower digital, single-modality threat models that dominate the benchmark literature this evaluation methodology largely grew out of.

8. Conclusion

The adversarial robustness literature reviewed here traces a specific arc: from the initial, somewhat unsettling discovery that state-of-the-art neural networks can be fooled by imperceptible, adversarially crafted perturbations [1,2], through an increasingly sophisticated attack methodology spanning white-box, black-box, and physical-world threat models [8,9,5,11,12,13], to formal verification and certified defense techniques that offer provable, rather than merely empirical, robustness guarantees for at least a subset of realistic network sizes and threat models [14,15,16,17]. Applied specifically to safety-critical domains — autonomous driving perception, aviation collision-avoidance, and clinical decision support — this literature demonstrates both that the threat is real and physically realizable, not merely a laboratory curiosity confined to digital pixel perturbations, and that the tools to evaluate and, in some cases, formally certify robustness against it already exist, if not yet at the scale every production system would require [19,20].

The methodological lesson that cuts across nearly every section of this synthesis is arguably more important than any single attack or defense result: a reported robustness number is only as credible as the attack used to produce it, and the field's own history — most starkly in Athalye et al.'s demonstration that the majority of a cohort of published defenses relied on gradient obfuscation rather than genuine robustness — shows how easily an under-evaluated defense can produce a robustness claim that collapses under closer scrutiny [3]. For a safety-critical systems engineer, the practical implication is that adversarial robustness evaluation cannot be a checkbox exercise performed once against a fixed, off-the-shelf attack suite; it needs to be treated as an adversarial, adaptive, and ideally formally-grounded

process, integrated into the same certification and safety-case discipline that already governs the rest of a safety-critical system's verification and validation lifecycle, and revisited as both attack techniques and the deployed model itself continue to evolve.

9. References

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2014) Intriguing Properties of Neural Networks. International Conference on Learning Representations (ICLR). DOI: [10.48550/arXiv.1312.6199](https://doi.org/10.48550/arXiv.1312.6199)
2. Goodfellow, I.J., Shlens, J., Szegedy, C. (2015) Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations (ICLR). DOI: [10.48550/arXiv.1412.6572](https://doi.org/10.48550/arXiv.1412.6572)
3. Athalye, A., Carlini, N., Wagner, D. (2018) Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80: 274-283. DOI: [10.48550/arXiv.1802.00420](https://doi.org/10.48550/arXiv.1802.00420)
4. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016) Concrete Problems in AI Safety. arXiv preprint. DOI: [10.48550/arXiv.1606.06565](https://doi.org/10.48550/arXiv.1606.06565)
5. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A. (2017) Practical Black-Box Attacks Against Machine Learning. Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIACCS), 506-519. DOI: [10.1145/3052973.3053009](https://doi.org/10.1145/3052973.3053009)
6. Kurakin, A., Goodfellow, I., Bengio, S. (2017) Adversarial Examples in the Physical World. International Conference on Learning Representations (ICLR) Workshop. DOI: [10.48550/arXiv.1607.02533](https://doi.org/10.48550/arXiv.1607.02533)
7. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P. (2016) DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2574-2582. DOI: [10.1109/CVPR.2016.282](https://doi.org/10.1109/CVPR.2016.282)
8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2018) Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations (ICLR). DOI: [10.48550/arXiv.1706.06083](https://doi.org/10.48550/arXiv.1706.06083)
9. Carlini, N., Wagner, D. (2017) Towards Evaluating the Robustness of Neural Networks. Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P), 39-57. DOI: [10.1109/SP.2017.49](https://doi.org/10.1109/SP.2017.49)
10. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A. (2016) The Limitations of Deep Learning in Adversarial Settings. Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 372-387.

DOI: [10.1109/EuroSP.2016.36](https://doi.org/10.1109/EuroSP.2016.36)

11. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D. (2018) Robust Physical-World Attacks on Deep Learning Visual Classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: [10.48550/arXiv.1707.08945](https://doi.org/10.48550/arXiv.1707.08945)
12. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J. (2017) Adversarial Patch. arXiv preprint. DOI: [10.48550/arXiv.1712.09665](https://doi.org/10.48550/arXiv.1712.09665)
13. Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q.A., Fu, K., Mao, Z.M. (2019) Adversarial Sensor Attack on LiDAR-Based Perception in Autonomous Driving. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2267-2281. DOI: [10.1145/3319535.3339815](https://doi.org/10.1145/3319535.3339815)
14. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J. (2017) Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. Proceedings of the 29th International Conference on Computer Aided Verification (CAV), Lecture Notes in Computer Science 10426: 97-117. DOI: [10.1007/978-3-319-63387-9_5](https://doi.org/10.1007/978-3-319-63387-9_5)
15. Huang, X., Kwiatkowska, M., Wang, S., Wu, M. (2017) Safety Verification of Deep Neural Networks. Proceedings of the 29th International Conference on Computer Aided Verification (CAV), Lecture Notes in Computer Science 10426: 3-29. DOI: [10.1007/978-3-319-63387-9_1](https://doi.org/10.1007/978-3-319-63387-9_1)
16. Wong, E., Kolter, J.Z. (2018) Provable Defenses Against Adversarial Examples via the Convex Outer Adversarial Polytope. Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80: 5286-5295. DOI: [10.48550/arXiv.1711.00851](https://doi.org/10.48550/arXiv.1711.00851)
17. Cohen, J., Rosenfeld, E., Kolter, Z. (2019) Certified Adversarial Robustness via Randomized Smoothing. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97: 1310-1320. DOI: [10.48550/arXiv.1902.02918](https://doi.org/10.48550/arXiv.1902.02918)
18. Croce, F., Hein, M. (2020) Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR 119: 2206-2216. DOI: [10.48550/arXiv.2003.01690](https://doi.org/10.48550/arXiv.2003.01690)
19. Julian, K.D., Lopez, J., Brush, J.S., Owen, M.P., Kochenderfer, M.J. (2016) Policy Compression for Aircraft Collision Avoidance Systems. Proceedings of the 35th IEEE/AIAA Digital Avionics Systems Conference (DASC), 1-10. DOI: [10.1109/DASC.2016.7778091](https://doi.org/10.1109/DASC.2016.7778091)
20. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S. (2019) Adversarial Attacks on Medical Machine Learning. Science, 363(6433): 1287-1289. DOI: [10.1126/science.aaw4399](https://doi.org/10.1126/science.aaw4399)
21. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M. (2019) Theoretically Principled Trade-off between Robustness and Accuracy. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97: 7472-7482. DOI: [10.48550/arXiv.1901.08573](https://doi.org/10.48550/arXiv.1901.08573)
22. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P. (2018) Ensemble Adversarial Training: Attacks and Defenses. International Conference on Learning Representations (ICLR). DOI: [10.48550/arXiv.1705.07204](https://doi.org/10.48550/arXiv.1705.07204)
23. Xie, C., Wu, Y., van der Maaten, L., Yuille, A.L., He, K. (2019) Feature Denoising for Improving Adversarial Robustness. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 501-509. DOI: [10.48550/arXiv.1812.03411](https://doi.org/10.48550/arXiv.1812.03411)
24. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A. (2019) Adversarial Examples Are Not Bugs, They Are Features. Advances in Neural Information Processing Systems (NeurIPS) 32. DOI: [10.48550/arXiv.1905.02175](https://doi.org/10.48550/arXiv.1905.02175)
25. Biggio, B., Roli, F. (2018) Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition, 84: 317-331. DOI: [10.1016/j.patcog.2018.07.023](https://doi.org/10.1016/j.patcog.2018.07.023)

26. Akhtar, N., Mian, A. (2018) Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6: 14410-14430. DOI: [10.1109/ACCESS.2018.2807385](https://doi.org/10.1109/ACCESS.2018.2807385)
27. Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K. (2020) Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing, 17(2): 151-178. DOI: [10.1007/s11633-019-1211-x](https://doi.org/10.1007/s11633-019-1211-x)

10. Declarations

Funding

This literature synthesis received no dedicated funding from public, commercial, or not-for-profit funding agencies.

Conflict of Interest

The authors declare that they have no known competing financial or personal interests.

Data Availability Statement

This article is a literature synthesis and did not generate new experimental data; all numerical results discussed are drawn from, and attributed to, the cited original publications, which should be consulted for their underlying data and code.

Ethical Approval

Not applicable. This work is a review of previously published literature and did not involve human participants, animal subjects, or new primary data collection.