International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Adversarial Attacks and Defense Mechanisms in Deep Neural Networks

Bhavesh Prajapati*, Bhavya Shah†

Submitted:13/05/2023 **Revised:** 28/06/2023 **Accepted:** 05/07/2023

Abstract—Deep neural networks (DNNs) achieve high predictive accuracy but remain vulnerable to deliberately engineered inputs and training-time manipulations. This paper presents a structured review of adversarial attacks and defense mechanisms. It organizes attacks by attacker knowledge, objective, perturbation model, interaction budget, and lifecycle stage; compares gradient-based, optimization-based, transfer, query-based, universal, physical, poisoning, and backdoor attacks; and evaluates defenses spanning adversarial training, preprocessing, detection, architectural regularization, verification, and certified robustness. Particular attention is given to evaluation failure modes such as gradient masking, non-adaptive testing, weak attack configurations, and mismatched threat models. The review finds that empirical robustness is highly conditional on the assumed perturbation set and evaluation procedure, while certified methods provide stronger guarantees but at a cost in scalability and clean accuracy. Adversarial training remains the most consistently effective empirical defense for norm-bounded image perturbations, yet it does not solve poisoning, physical-world, semantic, or cross-domain threats. The paper concludes with an evaluation protocol and a set of research directions for secure DNN deployment.

Index Terms—*adversarial examples, adversarial machine learning, deep neural networks, adversarial training, black-box attacks, poisoning, backdoors, certified robustness, model security, robustness evaluation.*

I. INTRODUCTION

Deep neural networks now support perception, language, speech, recommendation, cybersecurity, and autonomous decision pipelines, but their statistical success does not imply security under deliberate manipulation. The adversarial-example literature began with the observation that carefully chosen, often small, input perturbations can cause confident misclassification while leaving the semantic content apparently unchanged to a human observer. Subsequent work established stronger first-order attacks, optimization-based attacks, physical-world attacks, transfer attacks, data poisoning, and backdoor mechanisms. By 2023, adversarial machine learning had therefore become a lifecycle security problem rather than a narrow image-classification anomaly. NIST's March 2023 draft taxonomy explicitly treated evasion, poisoning, privacy, and related manipulations as security concerns across the machine-learning lifecycle [1]. Broad surveys likewise show a rapid expansion from image-space perturbations to graphs, text, audio, reinforcement learning, recommenders, and cybersecurity applications [2]-[10].

Two characteristics make adversarial robustness unusually difficult to measure. First, robustness is always relative to a threat model: the same model can be robust to an l-infinity perturbation budget yet fragile to l2, spatial, patch, semantic, query-based, or physical attacks. Second, the defender evaluates an adaptive opponent who can exploit differentiability, randomness, preprocessing, confidence outputs, or implementation details. Consequently, a defense that appears successful under one attack may fail under a stronger or better-adapted evaluation. Foundational studies on adversarial examples, projected-gradient adversaries, accuracy-robustness trade-offs, non-robust features, geometric attacks, and universal perturbations established much of this modern view [11]-[20].

This paper reviews adversarial attacks and defenses. The objective is not to claim a universally best defense, but to connect threat assumptions to appropriate attack families, defense mechanisms, and evaluation requirements. The central argument is that trustworthy robustness claims require three components simultaneously: a precisely stated

*IT Department, L. D. College of Engineering, Commissionerate of Technical

Education, Government of Gujarat, India

Email: b.b.prajapati@gmail.com, ORCID: 0000-0002-8015-7934

†Individual Scholar, Email: shahbhavya5800@gmail.com

attacker model, a defense designed for that model, and an evaluation that uses sufficiently strong adaptive attacks or formal certificates.

II. SCOPE, REVIEW METHOD, AND TERMINOLOGY

The review includes original research papers, peer-reviewed surveys, conference and journal publications, authoritative preprints, and the March 2023 NIST initial public draft. The selected literature covers test-time evasion, transferability, score- and decision-based black-box attacks, universal and physical perturbations, training-time poisoning and backdoors, cross-modal attacks, adversarial training, preprocessing, detection, architectural defenses, verification, and certified robustness. The paper emphasizes DNNs, while including graph, speech, text, video, and reinforcement-learning examples where they reveal general security principles.

An adversarial example is an input x_adv = x + delta selected to alter the prediction of a model f while satisfying constraints intended to preserve utility or perceptual similarity. For an untargeted attack, a common formulation is: minimize ||delta||_p subject to f(x + delta) != y and x + delta belonging to the valid input domain. For a targeted attack, the constraint becomes f(x + delta) = y_t. In practice, attackers often maximize a differentiable loss within a fixed perturbation set, while defenders optimize expected or worst-case loss over that set.

Attack taxonomies differ in naming, but the most operational dimensions are: attacker knowledge (white-box, gray-box, black-box), attacker objective (targeted or untargeted), lifecycle stage (training or inference), feedback (logits, scores, labels, or no queries), perturbation constraint (norm, sparsity, geometry, patch, semantic, or physical), and scope (per-example or universal). These dimensions are preferable to a single attack-name taxonomy because they transfer across modalities and model classes.

TABLE I
THREAT-MODEL DIMENSIONS FOR ADVERSARIAL DNN EVALUATION

Dimension	Representative choices	Why it matters
Knowledge	White-box; transfer; score-based; decision-based	Determines whether gradients, architecture, parameters, or only outputs are available.
Goal	Untargeted; targeted; confidence reduction	Changes the loss function and the success criterion.
Lifecycle	Training-time; inference-time	Separates poisoning/backdoors from evasion attacks.
Constraint	l0, l1, l2, l-infinity; spatial; patch; semantic; physical	Defines what constitutes an admissible adversarial modification.
Interaction	Zero-query transfer; bounded queries; unrestricted adaptive evaluation	Strongly affects practical black-box feasibility.
Scope	Instance-specific; universal; class-specific; trigger-based	Controls whether one perturbation generalizes across many samples.

III. ADVERSARIAL ATTACK MECHANISMS

A. Gradient-Based and Optimization-Based White-Box Attacks

White-box attacks assume access to model parameters and gradients. The fast gradient sign method uses the sign of the input gradient to construct a single-step perturbation, while projected gradient methods iterate first-order steps with projection back to the permitted norm ball. DeepFool approximates local decision boundaries, and Carlini-Wagner optimization directly searches for low-distortion adversarial inputs. These methods established strong baselines for security evaluation and showed that attack strength depends on optimization details, step sizes, restarts, loss functions, and stopping conditions [12], [13], [16], [17].

Spatial transformations, elastic-net objectives, momentum, and diverse parameterizations broadened the attack surface beyond a single norm. Universal perturbations demonstrated that a single perturbation can fool many inputs, undermining an intuitive assumption that adversarial errors are

idiosyncratic. These directions, together with sparse and structured perturbations, are represented in [21]-[25] and [20]-[23].

B. Black-Box, Transfer, Score-Based, and Decision-Based Attacks

Black-box attacks demonstrate that secrecy of model internals is not sufficient. Transfer attacks use a surrogate model and exploit the empirical transferability of adversarial examples, while score-based attacks estimate useful directions through output probabilities or logits. Zeroth-order optimization, AutoZOOM, bandit priors, and random-search attacks reduce the need for exact gradients. Decision-based attacks go further by requiring only the final class label. The literature from practical substitute-model attacks through Boundary Attack, ZOO, Square Attack, Sign-OPT, and SurFree forms a mature query-based threat family [26]-[37].

The security consequence is important: a deployed API can expose enough information for an adversary even when architecture and weights are hidden. Rate limits and score suppression may increase cost, but they do not eliminate transfer or label-only attacks. For black-box evaluation, robustness should therefore be reported against both transfer attacks and at least one strong query-based method, with the query budget stated explicitly.

C. Physical, Universal, Semantic, and Structured Attacks

Physical-world attacks must survive transformations introduced by printing, viewing angle, illumination, sensors, compression, distance, and temporal variation. Research has shown attacks against object detectors, vehicles, LiDAR perception, audio systems, and other sensing pipelines. The threat is qualitatively different from a purely digital norm-ball perturbation because the admissible transformation set is determined by the environment and sensor. Representative studies on robust real-world object attacks, camouflage, LiDAR, structured perturbations, differentiable rendering, and short-lived physical perturbations are included in [38]-[66].

A key lesson is that perceptual similarity is task dependent. A small l-infinity bound is convenient mathematically but may not correspond to semantic invariance. Spatial, patch, color, geometric, or physically realizable changes can be conspicuous in pixel norm while still preserving the underlying object identity. Robustness evaluation should therefore distinguish norm-bounded robustness from semantic and physical robustness rather than presenting a single scalar as universal security.

D. Cross-Modal and Task-Specific Attacks

Adversarial vulnerabilities appear beyond static image classification. Text classifiers can be fooled through discrete token or character changes; source-code models can be manipulated while preserving program behavior; audio attacks exploit temporal and spectral structure; video attacks exploit temporal redundancy; structured prediction can be attacked through task-level losses; and reinforcement-learning policies can be driven toward unsafe behavior. Representative examples are collected in [63]-[76]. The breadth of these results suggests that adversarial vulnerability is not tied to one architecture or one sensory modality.

E. Training-Time Poisoning and Backdoor Attacks

Poisoning attacks modify training data, labels, or the training process to degrade global accuracy or induce targeted behavior. Clean-label attacks are particularly concerning because poisoned examples can appear correctly labeled to human inspection. Backdoor attacks implant a trigger-response association so that ordinary inputs behave normally while triggered inputs are redirected. Work on back-gradient poisoning, generative poisoning, MetaPoison, clean-label "poison frogs," targeted backdoors, hidden triggers, and concentration-based limits illustrates the diversity of training-time threats [77]-[85].

Training-time attacks expose a different trust boundary from evasion attacks. Defending only the inference API is insufficient if the data pipeline, pretrained checkpoint, model supply chain, annotation process, or fine-tuning corpus can be influenced. Practical controls therefore include provenance, dataset versioning, access controls, anomaly screening, trigger testing, reproducible training, and independent evaluation of externally sourced models.

TABLE II
REPRESENTATIVE ADVERSARIAL ATTACK FAMILIES

Attack family	Typical access	Primary advantage	Representative literature

Attack family	Typical access	Primary advantage	Representative literature
FGSM / iterative first-order	White-box	Fast gradient exploitation and strong norm-bounded baselines	[12], [13]
DeepFool / C&W	White-box	Low-distortion optimization around decision boundaries	[16], [17]
Transfer attacks	Surrogate / black-box	Can require zero victim queries	[23], [26], [66]
ZOO / AutoZOOM / Square	Score-based black-box	Gradient-free or approximate-gradient optimization	[28], [29], [31]
Boundary / Sign-OPT / SurFree	Decision-based black-box	Works with class-label feedback only	[27], [33], [34]
Universal / physical	White- or black-box	Reusable perturbations and real-world relevance	[20], [38], [46], [67]
Poisoning / backdoor	Training access	Persistent behavior change or trigger insertion	[79], [82]-[84]

IV. DEFENSE MECHANISMS

A. Adversarial Training

Adversarial training minimizes loss on adversarially perturbed examples, commonly through a min-max objective: min_theta E_(x,y)[max_(delta in S) L(f_theta(x+delta), y)]. The inner maximization approximates the strongest attack within perturbation set S, and the outer minimization updates the model against those examples. Projected-gradient adversarial training remains one of the most consistently effective empirical defenses for norm-bounded image attacks [13], [109]. Ensemble, curriculum, spectral-normalization, domain-adaptation, and robust-distillation variants attempt to improve generalization or efficiency [109], [111]-[113], [115].

Important later refinements focus on the robustness-accuracy trade-off and training cost. TRADES provides an explicit objective balancing natural accuracy and boundary robustness [136]. Free and fast adversarial training reduce computational overhead [139], [140], while robust overfitting and catastrophic overfitting show that optimization efficiency can create new failure modes [140], [141]. MART emphasizes the treatment of misclassified examples during training [142]. These results make clear that adversarial training is not a single algorithm but a family whose robustness depends on the inner attack, budget, initialization, schedule, data, and early-stopping choices.

The main limitation is specificity. Training against one norm and radius does not guarantee robustness to a different norm, sparse perturbations, spatial transformations, patches, semantic changes, poisoning, or physical attacks. Strong adversarial training may also reduce clean accuracy, increase training cost, and require more data for robust generalization. Accordingly, claims should always state the exact perturbation set and evaluation attacks.

B. Input Transformation, Generative Purification, and Randomization

Preprocessing defenses attempt to remove adversarial structure before classification. Examples include input transformations, generative reconstruction, Defense-GAN, PixelDefend, stochastic activation pruning, and randomized transformations [93]-[95], [117], [118]. These methods can improve robustness to non-adaptive attacks, but preprocessing can be differentiable, approximated, or attacked through expectation over transformations. Thus, preprocessing should be treated as a component of a defense-in-depth design rather than assumed to create an impenetrable barrier.

Randomization is most compelling when accompanied by a formal certificate. Randomized smoothing converts a base classifier into a smoothed classifier with a probabilistic l2 robustness guarantee under Gaussian noise [137]. This distinction between heuristic randomness and certified smoothing is crucial: randomness alone may merely make gradients noisy, whereas a certificate establishes a radius within which the

prediction is provably unchanged under stated assumptions.

C. Detection, Rejection, and Interpretability-Based Defenses

Detection mechanisms attempt to flag adversarial inputs using confidence, latent features, intrinsic dimensionality, reconstruction error, attribution, random perturbations, or representation statistics. The literature includes face-recognition analyses, bypass studies, ML-LOO, local intrinsic dimensionality, subset scanning, and autoencoder-based methods [96], [97], [124], [126], [128], [129]. Interpretability and decision-boundary analyses can provide useful diagnostics [86], [125].

The central challenge is distribution shift under an adaptive attacker. A detector creates a new decision function, and the attacker can optimize against the combined classifier-plus-detector pipeline. Detection is therefore most useful when it has a clearly modeled rejection policy, low false-positive cost, and an adaptive evaluation that targets both evasion and detector bypass.

D. Architectural Regularization and Robust Representation Learning

Architectural defenses seek smoother or more stable functions through nonexpansive mappings, Bayesian modeling, spectral constraints, robust latent representations, and feature-level regularization. Representative studies include l2-nonexpansive networks, Adv-BNN, latent-layer defenses, spectral normalization, PeerNets, and graph-specific defenses [91], [107], [108], [112], [122], [123]. These approaches can complement adversarial training but do not by themselves establish universal robustness.

E. Verification and Certified Robustness

Certified robustness aims to prove that no adversarial example exists within a specified region. Early certified defenses and verification methods use convex relaxations, linear bounds, mixed-integer programming, ReLU stability, and abstract-domain or related bounding techniques [100]-[105]. These methods provide a fundamentally stronger type of claim than empirical attack resistance: within the certified radius and threat model, robustness does not depend on whether the evaluator happened to choose a sufficiently strong attack.

Certification introduces trade-offs. Exact methods often scale poorly with network size, while relaxations can become loose. Randomized smoothing improves scalability for l2 certificates [137], and PatchGuard illustrates task-specific certification against adversarial patches [119]. A practical secure system often combines empirical robustness testing with certificates where they are computationally feasible.

F. Domain-Specific Defenses

Defenses must respect modality structure. Audio defenses can exploit temporal consistency, while graph defenses reason about discrete topology and neighborhoods. WaveGuard and temporal-dependency analyses are examples in audio, and graph attack-defense work studies edge or topology manipulation [120], [121], [123]. The broader principle is that an effective defense should model the physically or semantically valid transformation set of its domain rather than inherit an image-specific norm constraint without justification.

TABLE III
DEFENSE FAMILIES, STRENGTHS, AND LIMITATIONS

Defense family	Strength	Main limitation	Representative references
Adversarial training	Strong empirical norm-bounded robustness	Cost; threat-model specificity; robust overfitting	[13], [136], [141]
Preprocessing / purification	Easy to compose with existing models	Often bypassed by adaptive gradients or EOT	[93]-[95], [138]
Detection / rejection	Can expose anomalous inputs	Detector itself becomes part of attack surface	[97], [124], [126]
Architectural regularization	Can improve smoothness and stability	Usually lacks universal guarantee	[91], [107], [112]
Formal certification	Provable robustness inside stated region	Scalability and looseness; narrow threat model	[100], [101], [104], [137]

Defense family	Strength	Main limitation	Representative references
Patch / modality-specific	Matches structured practical threats	Guarantees may not transfer across modalities	[119], [121], [123]

V. WHY DEFENSE EVALUATION FAILS

A. Gradient Masking and Obfuscated Gradients

A defense may appear robust because it makes optimization difficult rather than because it removes adversarial examples. Gradient masking can arise from non-differentiable operations, numerical instability, stochastic layers, saturation, or deliberate gradient obfuscation. Carlini and Wagner showed that numerous detector-based defenses could be bypassed [97], and later analysis systematized how obfuscated gradients produce false confidence [138]. A warning sign is when a weak one-step attack appears stronger than a multi-step attack, black-box attacks outperform white-box attacks, or robustness collapses when the defense is approximated by a differentiable surrogate.

B. Non-Adaptive Attacks

An adaptive attacker knows the defense mechanism and optimizes the full defended system. Evaluating only attacks designed for an undefended classifier is insufficient when preprocessing, randomization, detection, or rejection is present. Systematic adaptive studies have repeatedly broken defenses that initially appeared robust [143]. Adaptive evaluation may require expectation over transformation, backward-pass differentiable approximations, attacks on detector thresholds, multiple loss functions, or customized optimization. There is no single automatic adaptive attack that is guaranteed to test every defense correctly.

C. Weak Hyperparameters and Single-Attack Reporting

Attack results depend on step count, restarts, step size, confidence objective, initialization, precision, and norm projection. AutoAttack reduces evaluator degrees of freedom by combining complementary parameter-free attacks [24]. RobustBench similarly promotes standardized model and evaluation comparisons [144]. These tools do not eliminate the need for adaptive analysis, but they reduce accidental overestimation of robustness and improve reproducibility.

D. Mismatched Threat Models

A common reporting error is to compare defenses under different perturbation radii, architectures, preprocessing, data augmentation, or clean-accuracy regimes. Another is to interpret robustness to one l_p norm as security against unrestricted manipulation. Results should therefore report clean accuracy, robust accuracy, perturbation type and radius, attack knowledge, query budget, restarts, targeted/untargeted objective, and whether the attack is adaptive to the complete defense.

TABLE IV
MINIMUM EVALUATION CHECKLIST FOR ADVERSARIAL ROBUSTNESS

Item	Required reporting
Threat model	Attacker knowledge, goal, lifecycle stage, perturbation set, and query budget.
Clean baseline	Clean test accuracy and any preprocessing or rejection behavior.
White-box attacks	At least one strong iterative attack with restarts; defense-aware loss where applicable.
Black-box attacks	Transfer plus score- or decision-based attack when deployment exposes an API.
Adaptive testing	Attack the full pipeline, including randomization, detectors, preprocessing, and ensembles.
Cross-norm / semantic checks	Test plausible non-norm or alternate-norm threats when relevant to the application.

Item	Required reporting
Certification	Report certified accuracy and radius separately from empirical robust accuracy.
Reproducibility	State model, dataset, preprocessing, code/version, attack hyperparameters, and random seeds.

VI. COMPARATIVE ANALYSIS

The attack literature [21]-[76] demonstrates that adversarial capability is multidimensional. White-box attacks provide a strong local optimization test; transfer and black-box attacks model restricted access; physical and semantic attacks model deployment conditions; poisoning and backdoors attack the training lifecycle rather than the inference boundary. No single benchmark spans all of these dimensions. A model that performs well on one norm-bounded benchmark should therefore be described as robust under that benchmark, not "secure" without qualification.

The defense literature [86]-[135], complemented by the robust-training and evaluation works [136]-[144], suggests a hierarchy of evidence. At the lowest level is resistance to a small set of non-adaptive attacks. Stronger evidence comes from diverse white-box and black-box attacks, adaptive evaluation, standardized suites, and cross-threat testing. The strongest local evidence is a formal certificate, but certificates remain tied to a specified perturbation set and model assumptions. Security engineering should therefore combine these evidence types rather than substitute one for another.

Empirical adversarial training is best understood as robust optimization over a chosen uncertainty set. This explains both its strength and its limitations. It directly optimizes performance against a particular family of perturbations, making it unusually effective when the deployment threat resembles the training adversary. However, its guarantees do not automatically extend to sparse, geometric, semantic, patch, poisoning, or physical manipulations. Certified methods narrow this gap for specific threat models, while data provenance and lifecycle controls address threats that input-space robustness cannot solve.

VII. DEPLOYMENT GUIDANCE FOR SECURE DNN SYSTEMS

A secure deployment should begin with threat modeling, not with selecting a fashionable defense. The system owner should identify who can influence training data, whether models or checkpoints come from third parties, what inference outputs are exposed, what latency limits permit, and what manipulations are physically plausible. A perception model in an autonomous system needs physical and sensor-level testing; a public classification API needs query-based testing and rate controls; a fine-tuned enterprise model needs data-provenance and backdoor controls.

Second, robustness should be layered. Adversarial training can harden the classifier for an expected norm-bounded threat; input validation and provenance protect the training pipeline; detectors can provide monitoring or fail-safe behavior; rate limits can raise the cost of query attacks; and certification can provide high-assurance guarantees for carefully defined perturbation sets. The layers should be tested together because composition can introduce new gradients, thresholds, and failure modes.

Third, operational monitoring should distinguish security failures from ordinary distribution shift. Adversarial manipulation, corruption, sensor drift, and novel but benign inputs can all reduce confidence or alter internal representations. Logging should preserve enough context to support incident analysis while respecting privacy requirements. Model updates should trigger regression testing against the same adversarial suite plus newly relevant attacks, because robustness is not a one-time property.

VIII. OPEN RESEARCH CHALLENGES

The first challenge is threat-model realism. Norm balls are mathematically convenient, but the set of meaning-preserving changes is usually task dependent. Research must better connect perceptual, semantic, causal, and physical constraints to measurable robustness. This is particularly important for multimodal systems, embodied agents, and structured prediction.

The second challenge is scalable certification. Exact verification is expensive and relaxed bounds can be loose on large networks. Randomized smoothing scales better but primarily addresses particular norm settings. A major research direction is to obtain useful certificates for modern high-capacity models, richer transformations, patches,

and structured inputs without unacceptable accuracy or computation penalties.

The third challenge is robust generalization. Robust models often require more data and can exhibit robust overfitting. Pretraining, unlabeled data, augmentation, and representation learning can help, but the interaction between sample complexity, non-robust features, model capacity, and threat-model coverage remains unresolved. Results on robust generalization and pretraining highlight this tension [131], [132], [141].

The fourth challenge is evaluation standardization without complacency. Suites such as AutoAttack and RobustBench improve reproducibility, yet adaptive evaluation remains necessary for defenses with custom preprocessing, detectors, stochastic components, or novel threat models. Benchmarks should reduce accidental errors while preserving the attacker mindset required to test new defenses.

Finally, lifecycle security requires integrating adversarial robustness with poisoning, backdoors, privacy, software supply-chain security, and governance. The 2023 NIST draft reflects this broader framing [1]. For real systems, model security cannot be reduced to a single robust-accuracy number.

IX. CONCLUSION

Adversarial attacks reveal that high test-set accuracy is not equivalent to secure behavior. From first-order image perturbations to black-box queries, universal perturbations, physical attacks, poisoning, and backdoors, the literature shows that DNN vulnerabilities depend on attacker access and on the constraints used to define valid manipulation. Defenses are similarly conditional: adversarial training is a strong empirical baseline for a defined perturbation set, preprocessing and detection require adaptive evaluation, and certified methods trade flexibility and scalability for provable guarantees. The most defensible robustness claims therefore combine explicit threat models, strong adaptive attacks, standardized evaluation, and formal certification where feasible. For deployment, robustness should be treated as one layer in a broader machine-learning security lifecycle that includes data provenance, model supply-chain controls, monitoring, and repeated regression testing.

REFERENCES

[1] A. Oprea and A. Vassilev, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST AI 100-2 E2023, Initial Public Draft, National Institute of Standards and Technology, Mar. 8, 2023, doi: 10.6028/NIST.AI.100-2e2023.ipd.

[2] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity," ACM Computing Surveys, vol. 55, no. 8, Art. no. 163, pp. 1-39, Dec. 2022, doi: 10.1145/3547330.

[3] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, "A survey on adversarial attacks and defences," CAAI Trans. Intell. Technol. 6, 1, 25–45, 2021.

[4] Alexandru Constantin Serban, Erik Poll, and Joost Visser, "Adversarial examples on object recognition: A comprehensive survey," ACM Comput. Surv. 53, 3, 66:1–66:38. doi: 10.1145/3398394, 2020.

[5] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," ACM Comput. Surv. 55, 1, 1–38, 2021.

[6] Xingwei Zhang, Xiaolong Zheng, and Wenji Mao, "Adversarial perturbation defense on deep neural networks," ACM Comput. Surv. 54, 8, 1–36, 2021.

[7] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," ACM Comput. Surv. 54, 5, 1–36, 2021.

[8] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," Int. J. Autom. Comput. 17, 2, 151–178. doi: 10. 1007/s11633-019-1211-x, 2020.

[9] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu, "Adversarial attacks and defenses in deep learning," Engineering 6, 3, 346–360. doi: 10.1016/j.eng.2019.12.012, 2020.

[10] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," arXiv:2012.10544, 2020.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in Proc. 2nd International Conference on Learning Representations, 2014.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in Proc. 3rd

International Conference on Learning Representations, 2015.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. 6th International Conference on Learning Representations, 2018.

[14] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," in Proc. 7th International Conference on Learning Representations, 2019.

[15] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," in Proc. Annual Conference on Neural Information Processing Systems, 2019.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[17] Nicholas Carlini and David A. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symposium on Security and Privacy. IEEE Computer Society, 39–57. doi: 10.1109/SP.2017.49, 2017.

[18] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami, "The limitations of deep learning in adversarial settings," in Proc. IEEE European Symposium on Security and Privacy, 2016.

[19] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," in Proc. 5th International Conference on Learning Representations, 2017.

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in Proc. IEEE Conference on Computer Vision and Pattern Recognition. 86–94, 2017.

[21] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song, "Spatially transformed adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[22] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," in Proc. 32nd AAAI Conference on Artificial Intelligence, 2018.

[23] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in Proc. IEEE Conference on Computer Vision and Pattern Recognition. 9185–9193, 2018.

[24] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in Proc. International Conference on Machine Learning. PMLR, 2206–2216, 2020.

[25] Francesco Croce and Matthias Hein, "Mind the box: l _1-APGD for sparse adversarial attacks on image classifiers," in Proc. International Conference on Machine Learning. PMLR, 2201–2211, 2021.

[26] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning," in Proc. ACM on Asia Conference on Computer and Communications Security, 2017.

[27] Wieland Brendel, Jonas Rauber, and Matthias Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in Proc. 6th International Conference on Learning Representations, 2018.

[28] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in Proc. 10th ACM Workshop on Artificial Intelligence and Security, 2017.

[29] Chun-Chen Tu, Pai-Shun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng, "AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[30] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information," in Proc. International Conference on Machine Learning. PMLR, 2137–2146, 2018.

[31] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in Proc. European Conference on Computer Vision. Springer, 484–501, 2020.

[32] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in Proc. 7th International Conference on Learning Representations, 2019.

[33] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh, "Sign-opt: A query-efficient hard-label adversarial attack," arXiv:1909.10773, 2019.

[34] Thibault Maho, Teddy Furon, and Erwan Le Merrer, "SurFree: A fast surrogate-free black-box attack," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10430–10439, 2021.

[35] Andrew Ilyas, Logan Engstrom, and Aleksander Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," in Proc. 7th International Conference on Learning Representations, 2019.

[36] Nina Narodytska and Shiva Prasad Kasiviswanathan, "Simple black-box adversarial perturbations for deep networks," arXiv:1612.06299, 2016.

[37] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," IEEE Trans. Evolut. Computat. 23, 5, 828–841, 2019.

[38] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in Proc. the ACM SIGSAC Conference on Computer and Communications Security, 2019.

[39] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao, "Transferable adversarial attacks for image and video object detection," in Proc. 28th International Joint Conference on Artificial Intelligence, 2019.

[40] Chenxiao Zhao, P. Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen, "The adversarial attack and detection under the Fisher information metric," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[41] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang, "Heuristic black-box adversarial attacks on video recognition models," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[42] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong, "CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild," in Proc. 7th International Conference on Learning Representations, 2019.

[43] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z. Morley Mao, "Adversarial sensor attack on LiDAR-based perception in autonomous driving," in Proc. ACM SIGSAC Conference on Computer and Communications Security, 2019.

[44] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, and Patrick Le Callet, "A new ensemble adversarial attack powered by long-term gradient memories," in Proc. 34th AAAI Conference on Artificial Intelligence. 3405–3413, 2020.

[45] Rima Alaifari, Giovanni S. Alberti, and Tandri Gauksson, "ADef: An iterative algorithm to construct adversarial deformations," in Proc. 7th International Conference on Learning Representations, 2019.

[46] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, and Tadayoshi Kohno, "Physical adversarial examples for object detectors," in Proc. 12th USENIX Workshop on Offensive Technologies, 2018.

[47] Hiromu Yakura and Jun Sakuma, "Robust audio adversarial example for a physical attack," in Proc. 28th International Joint Conference on Artificial Intelligence. 5334–5341, 2019.

[48] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin, "Structured adversarial attack: Towards general implementation and better interpretability," in Proc. 7th International Conference on Learning Representations, 2019.

[49] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson, "Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer," in Proc. 7th International Conference on Learning Representations, 2019.

[50] Huan Zhang, Hongge Chen, Zhao Song, Duane S. Boning, Inderjit S. Dhillon, and Cho-Jui Hsieh, "The limitations of adversarial training and the blind-spot attack," in Proc. 7th International Conference on Learning Representations, 2019.

[51] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu, "A Frank-Wolfe framework for efficient and effective adversarial attacks," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[52] Tianhang Zheng, Changyou Chen, and Kui Ren, "Distributionally adversarial attack," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[53] Zhengli Zhao, Dheeru Dua, and Sameer Singh, "Generating natural adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[54] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song, "Generating adversarial examples with adversarial networks," in Proc. 27th International Joint Conference on Artificial Intelligence, 2018.

[55] Huy Phan, Yi Xie, Siyu Liao, Jie Chen, and Bo Yuan, "CAG: A real-time low-cost enhanced-robustness high-transferability content-aware adversarial attack generator," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[56] Kenneth T. Co, Luis Muñoz-González, Sixte de Maupeou, and Emil C. Lupu, "Procedural noise adversarial examples for black-box attacks on deep convolutional networks," in Proc. ACM SIGSAC Conference on Computer and Communications Security. 275–289, 2019.

[57] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon, "A survey on universal adversarial attack," in Proc. 30th International Joint Conference on Artificial Intelligence. ijcai.org, 4687– 4694. doi: 10.24963/ijcai.2021/635, 2021.

[58] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker, "On the effectiveness of low frequency perturbations," in Proc. 28th International Joint Conference on Artificial Intelligence. 3389–3396, 2019.

[59] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan, "Meta gradient adversarial attack," in Proc. IEEE/CVF International Conference on Computer Vision. 7748–7757, 2021.

[60] Xiaosen Wang and Kun He, "Enhancing the transferability of adversarial attacks through variance tuning," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1924–1933, 2021.

[61] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren, "Feature importance-aware transferable adversarial attacks," in Proc. IEEE/CVF International Conference on Computer Vision. 7639– 7648, 2021.

[62] Yan Feng, Bin Chen, Tao Dai, and Shu-Tao Xia, "Adversarial attack on deep product quantization network for image retrieval," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[63] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin, "Robust adversarial objects against deep learning models," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[64] Elias B. Khalil, Amrita Gupta, and Bistra Dilkina, "Combinatorial attacks on binarized neural networks," in Proc. 7th International Conference on Learning Representations, 2019.

[65] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, "Adversarial attacks on neural network policies," arXiv:1702.02284, 2017.

[66] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu, "Improving black-box adversarial attacks with a transfer-based prior," Adv. Neural Inf. Process. Syst. 32, 2019.

[67] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic, "SLAP: Improving physical adversarial examples with short-lived adversarial perturbations," in Proc. 30th USENIX Security Symposium (USENIX Security'21), 2021.

[68] Abdullah Hamdi, Matthias Mueller, and Bernard Ghanem, "SADA: Semantic adversarial diagnostic attacks for autonomous applications," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[69] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi, "Deep text classification can be fooled," in Proc. 27th International Joint Conference on Artificial Intelligence, 2018.

[70] Erwin Quiring, Alwin Maier, and Konrad Rieck, "Misleading authorship attribution of source code using adversarial learning," in Proc. 28th USENIX Security Symposium, 2019.

[71] Huangzhao Zhang, Zhuo Li, Ge Li, Lei Ma, Yang Liu, and Zhi Jin, "Generating adversarial examples for holding robustness of source code processing models," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[72] Xiaolei Liu, Kun Wan, Yufei Ding, Xiaosong Zhang, and Qingxin Zhu, "Weighted-sampling audio adversarial example attack," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[73] Hongting Zhang, Pan Zhou, Qiben Yan, and Xiao-Yang Liu, "Generating robust audio adversarial examples with temporal dependency," in Proc. 29th International Joint Conference on Artificial Intelligence, 2020.

[74] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su, "Sparse adversarial perturbations for videos," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[75] Yuan Gong, Boyang Li, Christian Poellabauer, and Yiyu Shi, "Real-time adversarial attacks," in Proc. 28th International Joint Conference on Artificial Intelligence. 4672–4680, 2019.

[76] Moustapha M. Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," Adv. Neural Inf. Process. Syst. 30, 2017.

[77] Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren, "Data poisoning attack against knowledge graph embedding," in Proc. 28th International

Joint Conference on Artificial Intelligence, 2019.

[78] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu, "Data poisoning against differentially-private learners: Attacks and defenses," in Proc. 28th International Joint Conference on Artificial Intelligence, 2019.

[79] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. 10th ACM Workshop on Artificial Intelligence and Security. 27–38, 2017.

[80] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen, "Generative poisoning attack method against neural networks," arXiv:1703.01340, 2017.

[81] W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein, "Metapoison: Practical general-purpose clean-label data poisoning," Adv. Neural Inf. Process. Syst. 33, 12080–12091, 2020.

[82] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," Adv. Neural Inf. Process. Syst. 31, 2018.

[83] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv:1712.05526, 2017.

[84] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash, "Hidden trigger backdoor attacks," in Proc. AAAI Conference on Artificial Intelligence, Vol. 34. 11957–11965, 2020.

[85] Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody, "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[86] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel, "Proper network interpretability helps adversarial robustness in classification," in Proc. International Conference on Machine Learning. PMLR, 1014–1023, 2020.

[87] Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang, "On the sensitivity of adversarial robustness to input data distributions," in Proc. 7th International Conference on Learning Representations, 2019.

[88] Tsui Wei, Huan Zhang, Pin Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho Jui Hsieh, and Luca Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in Proc. 6th International Conference on Learning Representations, 2018.

[89] Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska, "Global robustness evaluation of deep neural networks with provable guarantees for the Hamming distance," in Proc. 28th International Joint Conference on Artificial Intelligence. 5944–5952, 2019.

[90] Pengcheng Li, Jinfeng Yi, Bowen Zhou, and Lijun Zhang, "Improving the robustness of deep neural networks via adversarial training with triplet loss," in Proc. 28th International Joint Conference on Artificial Intelligence, 2019.

[91] Haifeng Qian and Mark N. Wegman, "L2-nonexpansive neural networks," in Proc. 7th International Conference on Learning Representations, 2019.

[92] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel, "Towards the first adversarially robust neural network model on MNIST," in Proc. 7th International Conference on Learning Representations, 2019.

[93] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[94] Pouya Samangouei, Maya Kabkab, and Rama Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in Proc. 6th International Conference on Learning Representations, 2018.

[95] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten, "Countering adversarial images using input transformations," in Proc. 6th International Conference on Learning Representations, 2018.

[96] Gaurav Goswami, Nalini K. Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in Proc. 32nd AAAI Conference on Artificial Intelligence, 2018.

[97] Nicholas Carlini and David A. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in Proc. 10th ACM Workshop on Artificial Intelligence and Security. 3–14, 2017.

[98] Ian Goodfellow, "Gradient masking causes clever to overestimate adversarial perturbation size," arXiv:1804.07870, 2018.

[99] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen, "Interpreting and evaluating neural network

robustness," in Proc. 28th International Joint Conference on Artificial Intelligence, 2019.

[100] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang, "Certified defenses against adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[101] Eric Wong and J. Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in Proc. 35th International Conference on Machine Learning, 2018.

[102] Aman Sinha, Hongseok Namkoong, and John C. Duchi, "Certifying some distributional robustness with principled adversarial training," in Proc. 6th International Conference on Learning Representations, 2018.

[103] Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, and Aleksander Madry, "Training for faster adversarial robustness verification via inducing ReLU stability," in Proc. 7th International Conference on Learning Representations, 2019.

[104] Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake, "Evaluating robustness of neural networks with mixed integer programming," in Proc. 7th International Conference on Learning Representations, 2019.

[105] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev, "Boosting robustness certification of neural networks," in Proc. 7th International Conference on Learning Representations, 2019.

[106] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian J. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[107] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh, "Adv-BNN: Improved adversarial defense through robust Bayesian neural network," in Proc. 7th International Conference on Learning Representations, 2019.

[108] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N. Balasubramanian, "Harnessing the vulnerability of latent layers in adversarially trained models," in Proc. 28th International Joint Conference on Artificial Intelligence. 2779–2785, 2019.

[109] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in Proc. 6th International Conference on Learning Representations, 2018.

[110] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay, "Cascade adversarial machine learning regularized with a unified embedding," in Proc. 6th International Conference on Learning Representations, 2018.

[111] Qi-Zhi Cai, Chang Liu, and Dawn Song, "Curriculum adversarial training," in Proc. 27th International Joint Conference on Artificial Intelligence. 3740–3747, 2018.

[112] Farzan Farnia, Jesse M. Zhang, and David Tse, "Generalizable adversarial training via spectral normalization," in Proc. 7th International Conference on Learning Representations, 2019.

[113] Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft, "Improving the generalization of adversarial training with domain adaptation," in Proc. 7th International Conference on Learning Representations, 2019.

[114] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proc. IEEE Symposium on Security and Privacy, 2016.

[115] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein, "Adversarially robust distillation," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[116] Jörn-Henrik Jacobsen, Jens Behrmann, Richard S. Zemel, and Matthias Bethge, "Excessive invariance causes adversarial vulnerability," in Proc. 7th International Conference on Learning Representations, 2019.

[117] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar, "Stochastic activation pruning for robust adversarial defense," in Proc. 6th International Conference on Learning Representations, 2018.

[118] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille, "Mitigating adversarial effects through randomization," in Proc. 6th International Conference on Learning Representations, 2018.

[119] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal, "PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking," in Proc. 30th USENIX Security Symposium (USENIX Security'21), 2021.

[120] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song, "Characterizing audio adversarial examples using temporal dependency," in Proc. 7th International Conference on Learning Representations, 2019.

[121] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar, "WaveGuard: Understanding and

mitigating audio adversarial examples," in Proc. 30th USENIX Security Symposium (USENIX Security'21), 2021.

[122] Jan Svoboda, Jonathan Masci, Federico Monti, Michael M. Bronstein, and Leonidas J. Guibas, "PeerNets: Exploiting peer wisdom against adversarial attacks," in Proc. 7th International Conference on Learning Representations, 2019.

[123] Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu, "Adversarial examples for graph data: Deep insights into attack and defense," in Proc. 28th International Joint Conference on Artificial Intelligence. 4816–4823, 2019.

[124] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan, "ML-LOO: Detecting adversarial examples with feature attribution," in Proc. 34th AAAI Conference on Artificial Intelligence, 2020.

[125] Warren He, Bo Li, and Dawn Song, "Decision boundary analysis of adversarial examples," in Proc. 6th International Conference on Learning Representations, 2018.

[126] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in Proc. 6th International Conference on Learning Representations, 2018.

[127] Bo Huang, Yi Wang, and Wei Wang, "Model-agnostic adversarial detection by random perturbations," in Proc. 28th International Joint Conference on Artificial Intelligence, 2019.

[128] Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, and Edward McFowland, "Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error," in Proc. 29th International Joint Conference on Artificial Intelligence. 876–882, 2020.

[129] Partha Ghosh, Arpan Losalka, and Michael J. Black, "Resisting adversarial attacks using Gaussian mixture variational autoencoders," in Proc. 33rd AAAI Conference on Artificial Intelligence, 2019.

[130] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8684–8694, 2020.

[131] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry, "Adversarially robust generalization requires more data," in Proc. 32nd International Conference on Neural Information Processing Systems. 5019–5031, 2018.

[132] Dan Hendrycks, Kimin Lee, and Mantas Mazeika, "Using pre-training can improve model robustness and uncertainty," in Proc. International Conference on Machine Learning. PMLR, 2712–2721, 2019.

[133] Adnan Siraj Rakin, Zhezhi He, Boqing Gong, and Deliang Fan, "Blind pre-processing: A robust defense method against adversarial examples," arXiv:1802.01549, 2018.

[134] Zeyuan Allen-Zhu and Yuanzhi Li, "Feature purification: How adversarial training performs robust deep learning," in Proc. IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS). IEEE, 977–988, 2022.

[135] Jing Wu, Mingyi Zhou, Ce Zhu, Yipeng Liu, Mehrtash Harandi, and Li Li, "Performance evaluation of adversarial attacks: Discrepancies and solutions," arXiv:2104.11103, 2021.

[136] 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. 36th Int. Conf. Machine Learning (ICML), PMLR, vol. 97, pp. 7472-7482, 2019.

[137] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in Proc. 36th Int. Conf. Machine Learning (ICML), PMLR, vol. 97, pp. 1310-1320, 2019.

[138] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in Proc. 35th Int. Conf. Machine Learning (ICML), PMLR, vol. 80, pp. 274-283, 2018.

[139] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial Training for Free!," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[140] E. Wong, L. Rice, and J. Z. Kolter, "Fast Is Better Than Free: Revisiting Adversarial Training," in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[141] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in Adversarially Robust Deep Learning," in Proc. 37th Int. Conf. Machine Learning (ICML), PMLR, vol. 119, pp. 8093-8104, 2020.

[142] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving Adversarial Robustness Requires Revisiting Misclassified Examples," in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[143] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, “On Adaptive Attacks to Adversarial Example Defenses,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[144] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, “RobustBench: A Standardized Adversarial Robustness Benchmark,” in NeurIPS Datasets and Benchmarks Track, 2021.