Journal of Automation and Intelligence 2 (2023) 175–197

Contents lists available at ScienceDirect

Journal of Automation and Intelligence

journal homepage: www.keaipublishing.com/en/journals/journal-of-automation-and-intelligence/

Review article

A comprehensive survey of robust deep learning in computer vision

Jia Liu [a], Yaochu Jin [b,*]

[a] Ping An Property & Casualty Insurance Company, Shenzhen, 518048, Guangdong, China
[b] School of Engineering, Westlake University, Hangzhou, 310030, China

ARTICLE INFO

Keywords:
Robustness
Deep learning
Computer vision
Survey
Adversarial attack
Adversarial defenses

ABSTRACT

Deep learning has presented remarkable progress in various tasks. Despite the excellent performance, deep learning models remain not robust, especially to well-designed adversarial examples, limiting deep learning models employed in security-critical applications. Therefore, how to improve the robustness of deep learning has attracted increasing attention from researchers. This paper investigates the progress on the threat of deep learning and the techniques that can enhance the model robustness in computer vision. Unlike previous relevant survey papers summarizing adversarial attacks and defense technologies, this paper also provides an overview of the general robustness of deep learning. Besides, this survey elaborates on the current robustness evaluation approaches, which require further exploration. This paper also reviews the recent literature on making deep learning models resistant to adversarial examples from an architectural perspective, which was rarely mentioned in previous surveys. Finally, interesting directions for future research are listed based on the reviewed literature. This survey is hoped to serve as the basis for future research in this topical field.

Contents

1. Introduction 177
2. Preliminaries 177
2.1. Perturbation and threat models 177
2.1.1. Perturbation 177
2.1.2. Threat models 178
2.2. Robustness 179
2.2.1. Strong robustness 179
2.2.2. Weak robustness 179
2.3. Evaluation metrics 180
2.3.1. Accuracy-based metrics 180
2.3.2. Network verification 180
3. Adversarial attacks 182
3.1. Digital attacks 182
3.1.1. White-box attacks 182
3.1.2. Black-box attacks 183
3.1.3. Gray-box attacks 185
3.2. Physical world attacks 185
4. Defense techniques 185
4.1. Detecting adversarial examples 186
4.2. Weakening adversarial examples 187
4.3. Strengthening models 187
4.4. Search for robust architectures 188
5. Open problems 188
6. Conclusion 189
CRediT authorship contribution statement 189
Declaration of competing interest 189
Data availability 190

[*] Corresponding author.
E-mail addresses: liujia578@pingan.com.cn (J. Liu), jinyaochu@westlake.edu.cn (Y. Jin).

https://doi.org/10.1016/j.jai.2023.10.002
Received 28 June 2023; Received in revised form 5 October 2023; Accepted 31 October 2023
Available online 20 November 2023

Acknowledgments 190
References 190

1. Introduction

Deep learning (DL) techniques have been demonstrated superior to conventional machine learning (ML) across various applications such as computer vision (CV) [1], natural language processing (NLP) [2], and speech recognition [3]. Due to the progress of DL recently, many breakthroughs in the field of CV have been made, including image classification [1,4–10], object detection [11,12] and semantic segmentation [13]. With the improvements of deep neural network (DNN) models, deep learning is rapidly evolving into safety-critical applications such as self-driving cars [14], video surveillance [15], drones and robotics [16]. Deep learning solutions, especially those derived from CV tasks, including facial recognition in ATMs [17] and Face ID security on mobile phones [18], play essential roles in our daily lives. Although DL performs many CV tasks with incredible accuracies, an intriguing weakness of DNNs in image classification was first discovered by Szegedy et al. [19]. When a DL model encounters deliberately designed adversarial images imperceptible to the human eyes, it is easy to be misled and make wrong predictions. Vision transformers [20], which have made state-of-the-art advances in classification tasks, also do not provide any improvements under white-box attacks [21]. If the adversarial robustness of visual models is not solved, it will also introduce security problems and risks to multimodal large language models (MLLM) since they integrate text and other modalities, especially vision. Mao et al. [22] showed that pre-trained large-scale visual language models like CLIP [23] have demonstrated that imperceptible adversarial perturbations can significantly reduce the model performance on new tasks, even though they generalize well on common unseen tasks. Therefore, robustness, the ability to deal with application errors or malicious inputs, is an essential requirement for DL.

Due to the security issues caused by adversarial examples, reviewing their rapid developments is of great significance. Although adversaries can deliberately conduct attacks on various tasks, this survey limits the research area to CV due to space constraints, the importance of visual tasks, and most publications aiming at visual tasks. There have been several surveys in adversarial DL [24–32], therein the works in [24,30,31] specifically focus on CV. The work mentioned above is either an overview of adversarial attacks and defenses against DL models or focuses on specific areas such as deep reinforcement learning and cybersecurity. This paper is different from theirs since we focus on not only adversarial DL but more generally robust DL. In addition, we provide a comprehensive review of robustness assessment metrics, which are essential for designing robust DNN models but have been less mentioned in previous surveys.

Fig. 1 overviews primary research topics of robust learning. It has been ten years since Szegedy et al. [19] discovered the vulnerability of DNNs under adversarial attack. This work summarizes publications on robust DL after Szegedy et al. [19], without paying attention to the uncertainties and security threats in various stages of ML systems that have existed for several decades. Altogether, this paper aims to make it an independent complete survey in the field by providing comprehensive information on robust DL in CV. It is hoped that through this survey, researchers can understand which research in the current domain is more valuable and has more practical significance instead of following the trend and inventing some unrealistic attack methods that cannot be realized.

The remainder of this paper is organized as follows. Section 2 is related to essential concepts and definitions to understand robustness properly. A summary of robustness evaluation measures is also presented in this section. Sections 3 and 4 review adversarial attacks and defense techniques, respectively. Section 5 outlines the main challenges and promising research paths of robust DL in CV. Finally, Section 6 brings the final considerations.

2. Preliminaries

This section provides basic concepts related to robust DL in CV. First, we introduce the essential components of model attacks and defenses in brief. Second, we clarify the definition of robustness from different perspectives. Third, we outline comprehensive robustness evaluation metrics, which are less discussed in previous surveys. We hope these preliminary studies help readers understand the critical research components related to robust DL.

2.1. Perturbation and threat models

2.1.1. Perturbation

Perturbation, or noise data, could be random or non-random noise added to original images. Fig. 2 shows examples of clean images and images with different types of random noise that are commonly found in digital images. These types of noise are often caused by poor picture-shooting environments and abnormal image sensors. Most random noise can be reduced through various image processing techniques, such as filtering and averaging. If the noise data are maliciously designed and constructed by adversaries to mislead the DL models, the perturbation is called adversarial perturbation, and the examples are adversarial examples.

Perturbation can be analyzed from the following perspectives [26]:

1. Perturbation scope. If the perturbation is generated for each clean image, such an attack is called an individual attack; if the perturbation can be applied to the entire data set, it is called a universal attack. While most attacks focus on generating adversarial examples on an individual attack basis, universal perturbations offer a more efficient method for deploying such attacks in the real world. With universal perturbations, deploying adversarial examples at scale is more accessible than individual attacks.
2. Perturbation limitation. The perturbation aims to minimize perturbation to make it imperceptible to humans. It can be achieved by setting perturbation as the objective function of an optimization problem. On the other hand, perturbation can also be set as a constraint in the optimization problem.
3. Perturbation measurement. The magnitude of perturbation can be measured by ℓ_p-norm:

$$\|x\|_p = \left(\sum_{i=1}^{n} \|x_i\|^p\right)^{\frac{1}{p}} \tag{1}$$

where ℓ_0, ℓ_2, and ℓ_∞ norm are three commonly used ℓ_p metrics. In the perturbed examples, the ℓ_0 norm counts the number of pixels. The ℓ_2 norm denotes the Euclidean distance between the perturbed and clean examples. The ℓ_∞ norm measures the maximum change for any pixels in the perturbed examples. However, ℓ_p norm does not match human perception well [33]. Perceptual adversarial similarity score (PASS) [34] is more consistent with human perception than ℓ_∞ norm. PASS values close to a pre-defined value τ indicate imperceptible perturbations, while smaller PASS values close to 0 suggest stronger modifications.

2.1.2. Threat models

Threat models in adversarial DL refer to the attacks that can be carried out against a DL model. According to how much information an adversary knows about the target model, adversarial attacks can be grouped into white-box, black-box, and gray-box attacks. We will elaborate on the prevalent adversarial attacks from this aspect in Section 3. Here we introduce adversarial attacks from three other perspectives, i.e., the attack phase, attack goal, and attack frequency.

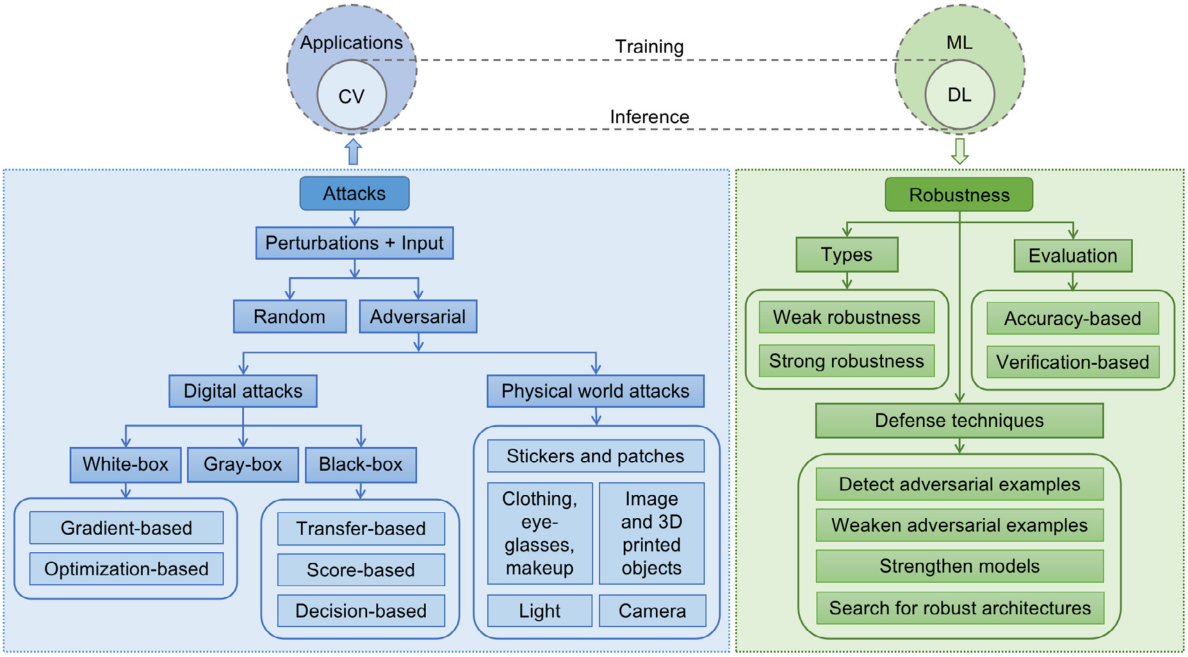

Fig. 1. Overview of research topics in robust learning. This paper only draws attention to the robustness of deep learning for computer vision tasks.

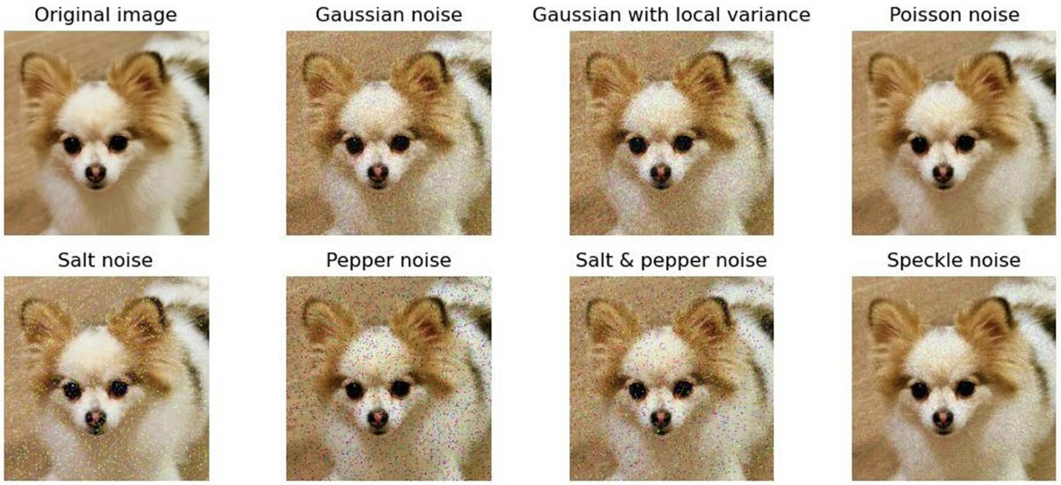

Fig. 2. Examples of different random noise. Original image: The noiseless image. Gaussian noise: Gaussian-distributed additive noise. Gaussian with local variance: Gaussian-distributed additive noise, with specified local variance at each point of the image. Poisson noise: Poisson-distributed noise generated from the data. Salt noise: Replace random pixels with 1. Pepper noise: Replace random pixels with 0. Salt & pepper noise: Replace random pixels with salt and pepper noise. Speckle noise: Multiplicative noise that the output $= x + x \cdot \mathcal{N}$, where x represents the original image and $\mathcal{N}$ is Gaussian noise with specified mean and variance.

1. Attack phase. According to the operation phases in which the adversary launches the attack, the attacks can be categorized into poisoning and evasion attacks. If adversaries can access the training database and insert fake or perturbed examples into the training database, the attacks are called poisoning attacks. These examples can lead to the trained classifier making wrong predictions on test samples [35]. If adversaries craft samples that the classifier fails to recognize correctly without manipulating the model or its parameters, it is called an evasion attack. Due to space limitations, this paper focuses on the attacks in the testing phase, a common assumption in adversarial DL.
2. Attack goal. The goal of targeted attacks is to make classifiers to label perturbed examples with specific labels. For instance, the adversarial images can disguise a face as an admin user [36]. Fig. 3 shows an example that a Maltese is misclassified as a tiger by the targeted attack. If the attacker just wants the classifier to make wrong predictions, and the victim sample has no specific target, such an attack is called an untargeted attack.
3. Attack frequency. One-step attacks require only a single step to craft adversarial examples, while iterative attacks take multiple steps to generate them. Iterative attacks typically generate stronger adversarial examples, but require more interactions with the victim classifier and require more computation time to generate. In cases where the task is computation-intensive, one-step attacks may be the only feasible choice. Therefore, when designing adversarial attacks, adversaries should consider the computational cost to prevent the attacks from becoming impractical in the real world.

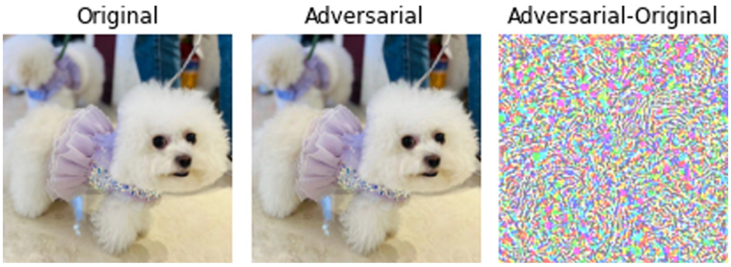

Fig. 3. An example of a targeted attack. A Maltese is misclassified as a tiger.

2.2. Robustness

Although the robustness of DL models has been explored in various papers, most work skipped clarifying the definition of robustness and left it as "you will know it when you see it". We aim to present the current robustness definitions literature.

Robustness has various interpretations in different scenarios. In optimization, robust solutions perform well under some degree of uncertainty. The robustness of an optimal solution can usually be discussed from the following points of view [37]: (1) The optimal solution is insensitive to minor variations of the design variables; (2) The optimal solution is susceptible to minor variations of environmental settings. The search for robust optimal solutions is of utmost significance in real-world applications. In computer systems, robustness indicates the ability of a computer system to handle errors during execution and cope with erroneous inputs. In DL, robustness usually expresses the requirement that the network behaves smoothly, i.e., that small input perturbations or minor model modifications should not cause significant spikes in the output of DNNs. Robustness is a relative, rather than an absolute, measure of model performance. Three key elements [38] may affect the robustness of DL models: (1) input data, (2) design or optimization of the network, and (3) evaluation measures of performance.

Previous research has defined the robustness of DL models in CV from different perspectives. For the convenience of description, we propose to use the term *strong robustness* to represent the insensitivity under an adversarial environment and the term *weak robustness* to describe the stability of DNN models in the non-adversarial environment, including noisy data.

2.2.1. Strong robustness

Focusing on the critical property of DNNs being invariant to perturbations at a given point (image), Bastani et al. [39] formalized pointwise robustness from [19,40,41]. Similarly, Katz et al. [42] used δ-local adversarial robustness to measure the resilience of a network against adversarial perturbations to specific inputs. More formally:

Definition 1. Network N is δ-locally-robust at point x_0 iff

$$\forall x'. \quad \left\| x' - x_0 \right\| \leq \delta \Rightarrow \arg\max N(x') = \arg\max N\left(x_0\right)$$

where N is a classifier associated with a set of labels L, x_0 is the original input, x' is the perturbed input that is close to x_0, "local" refers to a local neighborhood around x_0. Definition 1 states that for x', the network assigns to x' the same label that it assigns to x_0. Larger values of δ imply larger neighborhoods and hence better robustness. However, local robustness is checked for individual input points in an infinite input space without carrying over to other points that are not checked. Besides, for each x_0, the minimal acceptable value of δ should be specified, and these values vary between different input points.

In [42], Katz et al. also proposed an alternative approach, using the notion of global robustness, to overcome the need to specify each δ separately.

Definition 2. Network N is (δ, ϵ)-globally-robust in input region D iff

$$\forall x_1, x_2 \in D. \quad \left\| x_1 - x_2 \right\| \leq \delta \quad \Rightarrow \quad \forall l \in L. \quad \left| C\left(N, x_1, l\right) - C\left(N, x_2, l\right) \right| < \epsilon$$

where C denotes the confidence of model N that input is labeled l.

Compared with Definition 1, Definition 2 considers an input domain D instead of a specific point x_0, allowing it to cover infinitely multi-points or the entire input space in a single query, with δ and ϵ defined once for the entire domain. Moreover, it is more suitable to handle input points on the boundary between two labels. However, (δ, ϵ)-global robustness is significantly harder to prove on larger networks.

To balance the local and global robustness, Katz et al. [43] further proposed a hybrid definition:

Definition 3. Network N is (δ, ϵ)-locally-robust at point x_0 iff

$$\forall x'. \quad \left\| x' - x_0 \right\| \leq \delta \Rightarrow \forall l \in L. \quad \left| C(N, x', l) - C\left(N, x_0, l\right) \right| < \epsilon$$

Based on Definition 3, Katz et al. introduced a novel procedure Reluplex [43] to measure robustness, which will be introduced in Section 2.3.2.

2.2.2. Weak robustness

Strong robustness is geared towards malicious attackers, which focuses on the adversarial inputs and implicitly assumes that the attacker will succeed in finding them if such input exists. Weak robustness concerns random settings, where failures can occur naturally, not maliciously. This setup is more realistic for extensive systems expected to run at scale and is more likely to encounter random distortions than those crafted by malicious adversaries.

Mangal et al. [44] proposed probabilistic robustness, which requires neural networks to be robust to input distributions with at least $(1 - \epsilon)$ probability. Similarly, Levy and Katz [45] proposed the probabilistic robustness as follows:

Definition 4. The (δ, ϵ)-probabilistic-local-robustness score of a DNN N at input point x_0 is defined as:

$$\begin{aligned} \mathrm{Plr}_{\delta,\epsilon}\left(N, x_0\right) \triangleq\ & 1 - P_{x:\left\|x - x_0\right\| \leq \delta} \\ & \times \left[\left(\arg\max(N(x)) = \arg\max\left(N\left(x_0\right)\right) \vee C(x) < \epsilon\right)\right], \end{aligned}$$

where Plr is short for probabilistic-local-robustness, and it is a scalar value. Based on this definition, Levy and Katz [45] proposed robustness measurement and assessment (RoMA) to measure the model robustness to randomly-produced inputs.

As seen from the above, the definition of robustness is closely related to the evaluation of robustness. In the following, we will introduce the robustness evaluation methods.

2.3. Evaluation metrics

Research on DL in adversarial environments has focused mainly on different attack and defense methods. A reliable and quantitative measure of robustness is an essential prerequisite for the widespread use of neural networks in safety-related domains. However, few survey papers have summarized how to evaluate the robustness of DL models.

Guo et al. [46] summarized the robustness evaluation metrics into data-oriented and model-oriented metrics, which contain 23 evaluation metrics in total. Data-oriented metrics focus on whether the conducted evaluation covers most of the neurons within a model. In contrast, model-oriented metrics consider model behaviors and structures in the adversarial setting. In this survey, we focus on model-oriented evaluation metrics and improve the categories in [46]. We first give a summary of accuracy-based metrics that can be assessed by the model behaviors, followed by a description of main network verification approaches measured by model structures.

2.3.1. Accuracy-based metrics

The most straightforward approach to robustness evaluation is to use accuracy-based methods.

- Clean accuracy. Clean accuracy refers to the percentage of clean examples successfully classified by a classifier into the ground truth classes. If the model performs well in adversarial settings but poorly on clean images, it cannot be used in real scenarios.
- Attack success rate. The attack success rate indicates the proportion of misclassified adversarial examples for non-targeted attacks or represents the percentage of the adversarial examples classified as the target class for targeted attacks. A higher success rate indicates that the model is weaker and, thus, less resilient.
- Distortion. Leveraging ℓ_p norms, the distortion between clean and perturbed images can be measured. It usually suggests a robust model if it can resist a higher distortion.
- Robustness curves. Since models perform differently against the same attack under different attack parameters, Dong et al. [47] established a comprehensive benchmark to evaluate adversarial robustness using robustness curves, “accuracy vs. perturbation budget curve” and “accuracy vs. attack strength curve”.
- Robustness score under different attacks. To measure the resistance of model $\mathbf{A}$ under different types of attacks, Chang et al. [48] first calculated the difference between the attack accuracy of model m and the average attack accuracy of each attack executed on all models. Then the deviation was calculated to indicate the weight of each attack. Finally, the production of the deviation and the differences were used to evaluate the robustness.

2.3.2. Network verification

Network verification methods evaluate the robustness of DNNs by providing theoretically proven robustness bounds under specified perturbation constraints. Many scholars realized the importance of network robustness verification and introduced a variety of quantitative approaches to analyzing model robustness. Liu et al. [49] described various algorithmic approaches for verifying robustness and classified them based on whether they draw insights from these three categories of analysis: reachability, optimization, and search. Ji et al. [50] and Li et al. [51] provided a taxonomy for robustness verification from the perspectives of complete and incomplete approaches. The advantages and disadvantages of different network verification approaches are listed in Table 1, as discussed in [50].

1. Complete approaches. Based on the discrete optimization theory, complete methods aim to formally verify the feasibility of certain properties in neural networks for any possible input, using satisfaction modulo theory (SMT) or mixed integer linear programming (MILP). These methods typically achieve this by exploiting the piecewise linear properties of rectified linear units (ReLUs) and attempting to gradually satisfy the constraints they impose while searching for a feasible solution. For example, Katz et al. [42] proposed Reluplex, an SMT solver short for “ReLU with Simplex”, by extending the simplex algorithm to effectively tackle networks with ReLUs or max-pooling layers. Reluplex takes at least several hours to compute one sample for a small-scale network of about 100 neurons. Therefore, although Reluplex can formally verify some properties of neural networks, it cannot be extended to practical models due to the high computational cost. Furthermore, Reluplex analyzed the ReLU network by describing it as a piecewise linear function using a theoretical model in which matrix multiplications are linear. Cheng et al. [52] formalized calculating the model’s robustness boundary as a mixed integer programming (MIP) problem and designed a series of heuristic algorithms to encode the network function to reduce the MIP solver running time significantly. Specifically, they utilize a MIP solver based on the branch-and-bound algorithm to handle the encoding of nonlinear functions with integer variables, using a variant of big-M encoding strategy [53] to linearize nonlinear expressions. Furthermore, they defined a data flow analysis [54] that is used to generate relatively small big-M as the basis for accelerated solving MIP. Although MILP-based methods achieve promising results on small-scale networks, it is still a challenging problem to scale them to larger-scale networks. In addition, this method is only applicable to piecewise linear neural networks. To summarize, the complete method can prove the exact robustness boundary, but the computational complexity is high. In the worst case, the computational complexity exponentially increases relative to the network size, so it is usually only suitable for small-scale neural networks. Readers can refer to [50,51] if they are interested in more progress on the complete methods.
2. Incomplete approaches.
   Incomplete methods are efficient and scalable to complex neural networks but only demonstrate approximate robustness bounds. The incomplete verification can be categorized into verification via convex relaxation, abstract interpretation, Lipschitz’s constant, randomized smoothing, interval-bound propagation, cybernetics-based, and probability-based approaches.
   - Verification via convex relaxation. Many scholars transformed robust verification into convex optimization problems based on the ideas of semi-definite programming, duality, or nonlinear random projection [55–61]. Since such methods are unsuitable for large-scale neural networks, they are beyond the scope of this survey. Readers can read the above references for further understanding.
   - Verification via abstract interpretation. Based on abstract interpretation, the robustness of tiny neural networks can be certificated. Pulina et al. [62] first applied abstract interpretation [54] to model robustness analysis, but their method was only successful on networks with only six neurons. To apply abstract explanations to the analysis of larger and more complex neural networks, AI^2 [63], DeepZ [64], DiffAI [65], DeepPoly [66], RefineZono [67], k-ReLU [68] were presented. This survey does not detail these methods and refers interested readers to Refs. [63–68].
   - Verification via Lipschitz’s constant. Szegedy et al. [19] first calculated the global Lipschitz constant for every layer and used its product to explain the robustness problem in neural networks. Further, Hein and Andriushchenko [69] gave the bounds of robust space using local Lipschitz continuity conditions to achieve a more precise boundary. DeepGO [70] translates the robust spatial analysis of neural networks into reachability problems and uses adaptive optimization methods to solve this reachability problem. Since the network and function are Lipschitz continuous, all values between the upper and lower bounds are reachable. For a given input dataset, the robust spatial analysis of the neural network can be translated into the lower and upper limits of the Lipschitz continuous function that

Table 1
The advantages and disadvantages of different network verification approaches [50].

	Verification approaches	Advantages	Disadvantages
Complete	SMT, MILP	Prove the exact robustness boundary	Mainly support feed-forward networks on small-scale datasets
Incomplete	Convex relaxation	The efficiency is improved and can be applied to slightly larger neural networks compared to complete methods	Not suitable for large neural networks
	Abstract interpretation	The precision is relatively high; handle nonlinear activation functions and models with residual structures	The precision decreases as the number of layers of the neural network increases
	Lipschitz's constant	Can be applied to a wider range of neural networks	Looser robust bounds
	Randomized smoothing	Suitable for any DNNs	The decision boundary narrows with the adoption of the Randomized smoothing prediction rule; Enhanced perturbations do not necessarily solve the problem of boundary contraction, or even create other problems
	Interval bound propagation	Efficient and scalable	Unstable
	Cybernetics	The precision is greatly improved compared to methods based on the Lipschitz's constant,	Limited norms are applicable
	Probabilistic	More likely to be satisfied and verified with a looser definition of robustness	Assumptions about noise distribution are difficult to meet in practice

calculates its output value. However, they only focused on the feedforward neural networks. To obtain the robust boundary of the multi-layer complex neural network model, Weng et al. [71] proposed Fast-Lin and Fast-Lip. This approach is faster but only suitable for neural networks with ReLU. To apply neural network models with different structures, Weng et al. [72] adopted extreme value theory and proposed **C**ross **L**ipschitz **E**xtreme **V**alue for n**E**twork **R**obustness (CLEVER). It is a generic attack-agnostic metric and computationally feasible for large neural networks. Subsequently, Weng et al. [73] proposed two extensions to CLEVER. Additionally, based on the relaxation of the polynomial optimization problem [74], Latorre et al. [75] proposed a general method, dubbed LiPopt, to calculate the upper limit of the Lipschitz constant of the neural network. The LiPopt leverages the polynomial inequality to describe the unit ball, thus covering both the ℓ_2 norm and the ℓ_∞ norm.

- Verification via randomized smoothing. Although researchers have proposed many robust boundary analysis methods for models, these methods rarely scale to large neural networks trained on challenging datasets such as ImageNet, and the variety of models applicable is limited. To address this issue, PixelDP [76] first proposes the concept of "randomized smoothing" by using the link between differential privacy and model robustness, providing a robust guarantee for random smoothing of top-1 predictions. To obtain a more accurate robustness boundary, Cohen et al. [77] used the Neyman–Pearson lemma to obtain a verifiable range of ℓ_p ball radius under the Gaussian random noise assumption. Pinot et al. [78] further theoretically proved the effectiveness of the randomized smoothing and gave the robustness boundary of the stochastic smoothing model when the noise obeys the exponential distribution. Lee et al. [79] extended the distribution types of perturbations from the ℓ_2 norm continuous space assumption in [77] to the ℓ_0 norm discrete space assumption. In addition, the researchers also made a preliminary discussion on the robustness boundary of the model under the assumption that the noise obeys the uniform distribution [80] and the multinomial distribution [81]. Moreover, Salman et al. [82] designed an adaptive attack on stochastic smooth classifiers and used this attack in an adversarial training setting to enhance the provable robustness of smooth classifiers. In subsequent work, Dvijotham et al. [83] extended this idea to arbitrary smoothing methods, using f-divergence to demonstrate the robustness of smoothing classifiers. Additionally, Salman et al. [84] adopted a black-box randomized smoothing method for the pre-trained model by adding a denoiser before the existing image classification model and using random smoothing to obtain a new classification model. Without modifying the pre-trained classification model, it guarantees ℓ_p robustness to adversarial examples. Jia et al. [85] considered the robustness of top-k predictions, which are more widely used in practice. Wang et al. [86] and Weber et al. [87] studied the robustness of the model to backdoor attacks based on randomized smoothing. Different from the above research, Mohapatra et al. [88] first pointed out the side effects of current stochastic smoothing methods: (1) the decision boundary shrinks with the adoption of stochastic smoothing prediction rules; (2) enhancing the noise does not necessarily solve the problem of boundary shrinkage, and even creates other problems.
- Verification via interval bound propagation. ReluVal [89] verifies robustness by using symbolic interval propagation. However, the computational cost is high, so it is only suitable for small-scale simple neural networks with only a few hidden layers, and it is difficult to scale to complex, large-scale neural networks in practice. In subsequent work, Wang et al. [90] further optimized the relaxation method for the ReLU nonlinear activation function and obtained a tighter model robustness boundary. Based on interval bound propagation (IBP) [91], Gowal et al. [92] proposed a verifiable robust network training method, in which IBP defines a loss function to minimize the upper bound of the maximum difference between any logarithm when the input data is perturbed "within the sphere limited by the ℓ_∞ norm". This method is fast compared to [55, 65,93], which makes this approach highly scalable. Additionally, Zhang et al. [94] explored a new AT method CROWN-IBP with robustness guarantee by combining the forward propagation IBP boundary [92] and the backward

propagation boundary. Based on compact linear relaxation CROWN [95], CROWN-IBP is computationally efficient and superior to the IBP method.
- Verification via cybernetics-based approaches. A few researchers attempted to verify the model robustness through cybernetics-based approaches. Wang et al. [96] investigated combining static neural network verification tools with robust control theory to verify neural network robustness in control loops. Wang et al. [97] unified the optimal control theory of the transport equation with the training and testing practice of ResNets, and proposed a simple and effective ResNets ensemble algorithm to improve the accuracy of rigorously trained models on clean and adversarial images. Carr et al. [98] proposed a method for automatically extracting finite-state controllers from RNNs by integrating formalization methods and ML techniques, suitable for existing model robustness formal verification tools when composed with finite-state system models. Although the accuracy of cybernetics-based methods is improved compared to traditional Lipschitz-based methods, such methods have limited norm perturbations.
- Verification via probability-based approaches. Probability-based approaches can be employed to verify the weak robustness. Mangal et al. [44] employed abstract interpretation and importance sampling to verify whether the neural network is probabilistically robust. Concurrently, Weng et al. [99] proposed PROVEN to study the probabilistic robustness of the model when the perturbation follows a specific probability distribution and statistically provide a probability guarantee that the top-1 prediction result of the model will not change for arbitrary bounded LP perturbations. Assumed that the input uncertainty is random and infinite, Fazlyab et al. [100] abstracted nonlinear activation functions on their input–output pairs by imposing them on a combination of abstraction and quadratic constraints and then used SDP to analyze the security of the abstracted network satisfying the security of the original network, providing statistical guarantees for the output of the neural network. Web et al. [101] assessed the model robustness by estimating the proportion of inputs for which a property is violated. Specifically, they estimated the probability of the event that the property is violated under an input model. They demonstrated that this method could emulate formal verification procedures on benchmark problems while scaling to more extensive networks and providing reliable additional information in the form of accurate estimates of the violation probability. Short for Robustness Measurement and Assessment, RoMA [45] determines the probability that a random input perturbation might cause misclassification. RoMA focuses on a random setting where perturbations can occur naturally and are not necessarily malicious.

Besides, DeepSafe [102] leverages a data-guided methodology to determine safe regions, which characterizes the behavior of the network over partitions of the input space and makes the network amenable to analysis and verification. This technique could be coupled with any verification technique. To summarize, verifying the robustness of DNNs is a challenging task. Most work for verification of the robustness of DNNs suffers from scalability issues and can only be applied to small-scale or medium-scale DNNs. Full benchmark results and a leaderboard on robustness verification of DNNs are available on https://github.com/AI-secure/VeriGauge, which is developed by Li et al. [51] to assist in network verification.

In addition to accuracy-based metrics and network verification, Wang et al. [103] propose a robustness predictor by building an additional DNN model, taking the target model's penultimate layer as input and training it with robustness values for the target model's training data computed with adversarial searching. Besides, Carlini et al. [104] provided principles for performing defense evaluations and a checklist for avoiding evaluation pitfalls, which help researchers evaluate attack or defense methods comprehensively.

3. Adversarial attacks

Several papers have reviewed adversarial attacks on CV tasks from different perspectives. In 2018, Serban et al. [105] divided the attack methods into optimization-based, sensitive feature-based, geometric transformation-based, generative model-based, gray and black-box, and other attacks. Moreover, in 2020, Serban et al. [106] presented attack taxonomies according to the goals, knowledge, strategies, and performance of adversaries and classified existing approaches based on the taxonomies. Categorized by different visual tasks, Akhtar and Mian [24] reviewed the adversarial attacks for image classification and other tasks and presented some adversarial attacks in the real world. After this, Akhtar et al. [107] extended the original paper in the advances on CV and expanded it with more recent attacks. According to the phase in which attacks occur, Qiu et al. [108] described attacks on executing the training and testing phases of networks. In addition to the above classification methods, attack approaches can also be grouped into gradient-based, score-based, transfer-based, and decision-based attacks [109]. Due to the limitation of previous categorization methods, the newly presented attacks are difficult to classify. Ding and Xu [30] combined [105,109] and added functional-based attacks to the taxonomy. According to adversaries' knowledge and the attack scenarios, Xu et al. [110] reviewed white-box, black-box, gray-box, physical world attacks, and poisoning attacks. In 2021, Kong et al. [111] also introduced representative white-box, black-box, and adversarial attacks in the physical world. Similar to the categories in [109], Long et al. [112] summarized and analyzed the latest and representative attacks on CV based on gradient-based, transfer-based, score-based, and geometric-transformation-based attacks.

Based on the above review articles, this survey will classify adversarial attack methods according to the attack implementation scenarios: digital and physical world attacks. In the digital attack, according to the attacker's knowledge, the attacks are divided into white-box, black-box, and gray-box attacks.

3.1. Digital attacks

3.1.1. White-box attacks

White-box attacks refer to attacks in which adversaries know the model information, including its architecture and internal parameters. Gradient-based attacks are the most common white-box attacks. Adversaries attempt to attack by computing the gradient of the model's loss function with respect to the input data and then modifying the input data in the direction that maximizes the loss. These attacks are very effective and have been shown to work on many DL models. Here is a brief introduction to representative attacks and the characteristics of white-box attacks are presented in Table 2.

- **L**imited-emory **B**royden **F**letcher **G**oldfarb **S**hanno algorithm (L-BFGS). Szegedy et al. [19] proposed L-BFGS attacking models by considering solving a box-constrained optimization problem. Although the L-BFGS method performs well, it is expensive to calculate adversarial examples.
- Fast gradient sign method (FGSM). This method is the earliest gradient-based adversarial attack, and many scholars have proposed some improvements based on FGSM in subsequent research. Proposed by Goodfellow et al. [40], FGSM crafts adversarial examples using small perturbations along the gradient to maximize the loss.

$$x' = x + \epsilon \operatorname{sign}\left(\nabla_x J_\theta(x, l)\right) \tag{2}$$

Table 2
A summary of the characteristics of white-box adversarial attacks.

Methods	Attack goal	Attack frequency	Perturbation scope	Perturbation norm
L-BFGS [19]	Targeted	Iterative	Individual	ℓ_∞
FGSM [40]	Targeted/untargeted	One-step	Individual	ℓ_∞
BIM [113]	Targeted/untargeted	Iterative	Individual	ℓ_∞
PGD [114]	Targeted/untargeted	Iterative	Individual	ℓ_∞
MI-FGSM/MIM [115]	Untargeted	Iterative	Individual	ℓ_∞
NI-FGSM [116]	Untargeted	Iterative	Individual	ℓ_∞
JSMA [117]	Targeted/untargeted	Iterative	Individual	ℓ_0
Deepfool [118]	Untargeted	Iterative	Individual	ℓ_2, ℓ_∞
UAP [118]	Untargeted	Iterative	Universal	ℓ_2, ℓ_∞
C&W [119]	Targeted/untargeted	Iterative	Individual	$\ell_0, \ell_2, \ell_\infty$
LogBarrier [120]	Untargeted	Iterative	Individual	ℓ_∞

where x and x' denote the clean example and adversarial example, respectively. The ϵ refers to the step size, l denotes the label of the data, θ denotes the model parameters, J represents the loss function, ∇_x refers to the gradient to x. The FGSM algorithm only requires one-step gradient updates, so it is efficient to attack the model. However, the one-step update may not be enough to attack the model successfully.

- Basic iterative method (BIM). Kurakin et al. [113] extended FGSM by iteratively taking multiple small gradient steps called I-FGSM or BIM.

$$\begin{aligned} x'_0 &= x \\ x'_{n+1} &= \mathrm{Clip}_{x,\xi}\left\{x'_n + \epsilon \operatorname{sign}\left(\nabla_x J\left(x'_n, l\right)\right)\right\} \end{aligned} \tag{3}$$

where Clip $_{x,\xi}\left\{x'\right\}$ limits the range of the generated examples in each iteration:

$$\mathrm{Clip}_{\ x,\xi}\left\{x'\right\} = \min\left\{255, x+\xi, \max\left\{0, x-\epsilon, x'\right\}\right\} \tag{4}$$

- Projected gradient descent (PGD). By initializing adversarial examples at random points within the allowed norm range and then running BIM for multiple iterations, a stronger gradient attack than BIM and FGSM is proposed by Madry et al. [114], called PGD.
- Momentum iterative FGSM (MI-FGSM or MIM). Dong et al. [115] integrated momentum term into the BIM and proposed MI-FGSM. This method helps escape local maxima during optimization, leading to a higher attack success rate and transferability than other attacks.
- Nesterov iterative FGSM (NI-FGSM). Adopting Nesterov accelerated gradient into the iterative attacks, the NI-FGSM [116] can improve the transferability of adversarial examples effectively.
- Jacobian-based saliency map attack (JSMA). Different from the above approaches based on FGSM, Papernot et al. [117] employed an efficient saliency adversarial map and proposed JSMA to conduct adversarial attacks.
- Deepfool. Employing the linearity assumption of the neural network to simplify the optimization process, Moosavi-Dezfooli et al. [118] proposed DeepFool that searches for the closest decision boundary and perturbs the input towards the decision boundary. DeepFool repeatedly perturbs the input until the model makes wrong predictions.
- Universal adversarial perturbations (UAP). Extended from Deepfool [118], Moosavi-Dezfooli et al. [121] proposed that UAP is a universal attack.
- C&W attack. Carlini and Wagner [119] introduced the C&W attack, optimizing the distances from clean examples and adversarial examples. They explored generating adversarial examples under ℓ_0, ℓ_2, and ℓ_∞ norm and seven modified objective functions. Fig. 4 presents an example of C&W attack that can mislead the classifier to recognize a toy poodle into a leopard. The C&W attack can also escape from defensive distillation [119].
- LogBarrier attack. Finlay et al. [120] designed a LogBarrier attack that employs training loss functions to mislead the model, wherein the logarithmic barrier method [122] is adopted to execute a non-targeted attack.

3.1.2. *Black-box attacks*

Black-box attacks refer to attacks where adversaries know little model information and rely solely on external observations to carry out the attack. According to the access and resource restrictions in real-world scenarios, the black-box settings can be summarized into query-limited, partial-information, and label-only settings. Therein, attacks under partial-information settings are also called gray-box attacks, which will be described in Section 3.1.3. Black-box attacks can be categorized into transfer-based, score-based, and decision-based black-box attacks. A category of black-box attacks is shown in Table 3.

1. Transfer-based attacks. Transfer-based attacks are black-box attacks with substitute models that make use of the transferability of adversarial examples and employ adversarial examples generated on substitute models to attack the target model. To achieve high transferability, the following approaches have been proposed.

   - Adding a momentum term. To improve the transferability, MI-FGSM [115] was proposed which is also a white-box attack as mentioned in Section 3.1.1. Instead of applying the perturbation directly to the image, a momentum term is added. The momentum term helps the attack to escape from local maxima, which can be particularly useful when attacking models with non-smooth decision boundaries.
   - Ensemble-based method. Liu et al. [123] hypothesized that if an adversarial example remains adversarial to multiple models, it might also transfer to others. They developed an ensemble approach whose basic idea is to generate adversarial images for model ensembling.
   - Translation-invariant (TI) method. To further improve the transferability, a TI method [124] was also designed, which first generates an initial adversarial example using a standard attack method, such as the FGSM. Then, the perturbation is applied to the Fourier transform of the image. This has the effect of making the perturbation invariant to translation since Fourier transforms are translation-invariant. Once the perturbation has been applied to the Fourier transform of the initial image, it can be applied to other images in the dataset.
   - Diverse inputs method (DIM). The DIM [125] generates multiple adversarial examples for a single input by optimizing the loss function with respect to a set of random initializations. Specifically, DIM first generates a set of random initializations for the adversarial perturbation. Then, the loss function is optimized for each initialization using an iterative optimization method, such as BIM or PGD. The resulting set of adversarial examples is combined into a single set of diverse adversarial examples using a diversity-promoting loss function, which encourages the set of adversarial examples to be diverse by penalizing the similarity between the examples.

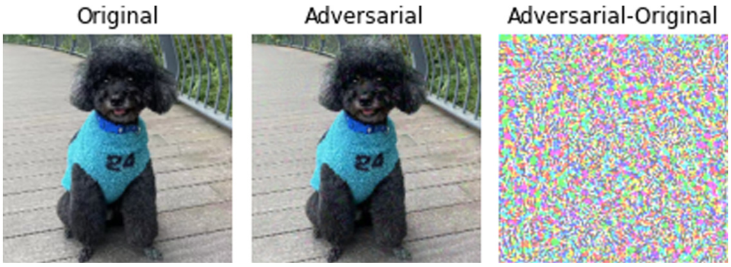

Fig. 4. An example of C&W attack. A toy poodle is misclassified as a leopard.

Table 3
A category of black-box adversarial attacks.

Black-box attacks	Methods	References
Transfer-based attacks	Adding a momentum term	[115]
	Ensemble-based method	[123]
	Translation-invariant method	[124]
	Diverse inputs method	[125]
	Intermediate level attack	[126]
	TREMBA	[127]
	Variance tuning based method	[128]
Score-based attacks	Zeroth Order Optimization	[129]
	Natural evolutional strategies	[130]
	Simultaneous perturbation stochastic approximation	[131]
	$\mathcal{N}$ATTACK	[132]
Decision-based attacks	The boundary attack	[133]
	Evolutionary attack	[134,135]
	IoU attack	[136]

- Intermediate level attack (ILA). Huang et al. [126] introduced ILA with two variants, ILAP and ILAF, where “P” and “F” refer to “projection” and “flexible”, respectively. Unlike other attacks that focus on perturbing the input data, ILA targets the internal representations of the model, making attack detection and defenses more difficult. In an ILA, a generative model is trained to learn the distribution of intermediate representations generated by the target model. Then, adversaries craft adversarial examples by sampling from this learned distribution and feeding the samples into the target model.
- **TR**ansferable **EM**bedding based **B**lack-box **A**ttack (TREMBA). Huang et al. [127] introduced TREMBA that generates transferable adversarial examples by learning a low-dimensional embedding using a pre-trained model. A search is then conducted within the embedding space to attack an unknown network.
- Variance tuning based method. To stabilize the direction of the update gradient, Wang and He [128] adjusted the current gradient to get rid of the local optimum, taking into account the gradient variance of the previous iteration. This method does not exist solely but can be combined with iterative gradient-based methods.

2. Score-based attacks. Score-based attacks are black-box attacks with gradient estimation, relying on querying the model to obtain its predicted scores for a given input. Adversaries then use these scores to generate adversarial examples. Score-based attacks aim to find a perturbation that maximizes the difference between the predicted score for the correct class and the predicted score for the target class. Even though the model gradients are unavailable, the gradient-free optimization techniques can help estimate them through queries.

    - Zeroth Order Optimization (ZOO). Chen et al. [129] proposed a derivative-free method, ZOO, estimating the gradient at each coordinate by finite differences and employing C&W for attacks. They leveraged zeroth order stochastic coordinate descent, dimension reduction, hierarchical attack, and importance sampling techniques to reduce the number of queries.
    - Natural evolutional strategies (NES). Ilyas et al. [130] employed a variant of NES [137] to estimate the full gradient and performed a PGD [114] with momentum based on the NES gradient estimate.
    - Simultaneous perturbation stochastic approximation (SPSA). Uesato et al. [131] approximated the gradients using SPSA [138].
    - $\mathcal{N}$ATTACK. Li et al. [132] proposed $\mathcal{N}$ATTACK, learning a Gaussian distribution centered around the input so that the samples drawn from this distribution are likely adversarial examples.

3. Decision-based attacks. In label-only settings, although decision-based attacks are more challenging since the model only provides the final decision of the model, they are the most realistic in real-world applications.

    - The boundary attack. The boundary attack [133] crafts adversarial examples by performing a local search in the vicinity of the decision boundary of a model. The attack starts with an initial random perturbation of the input. It iteratively moves the perturbation towards the decision boundary while ensuring that the perturbed input remains within a predefined ℓ_∞ distance from the original input.
    - Evolutionary attack. Based on differential evolution, Su et al. [134] estimated the location and RGB value of the pixel to be modified in the image to craft adversarial examples, demonstrating that DNNs can even be fooled by restricting the perturbation to a single pixel. Employing (1+1) covariance matrix adaptation evolution strategy (CMA-ES) [139], Dong et al. [135] proposed the evolutionary attack on face recognition systems. The algorithm starts with an initial population of candidate adversarial examples and iteratively evolves the population by selecting the most successful candidates and applying genetic operators

such as mutation and crossover to generate new candidates. The fitness of each candidate is evaluated based on its success rate against the target model. The algorithm terminates when an effective adversarial example is generated or a maximum iteration is reached.
- IoU attack. In the field of computer vision, most adversarial attack methods are aimed at image classification tasks, which are static. Target tracking is also one of the important tasks of computer vision. Jia et al. [136] proposed an IoU attack for object tracking. IoU is the intersection area of the predicted bounding box and the real bounding box divided by their union area. The IoU score will affect the prediction results of the next step of object tracking. The IoU attack calculates the weighted scores of two types of IoUs to find a gradually decreasing IoU score while introducing the lowest noise.

3.1.3. Gray-box attacks

Gray-box attacks, also named semi-white box attacks, merely requiring partial knowledge of the victim model, are often seen as a special case of the black box scenario.

1. Similarity-based gray-box adversarial attack (SGADV). Wang et al. [140] proposed SGADV by developing a novel objective function using similarity score to increase the attack performance against the face recognition authentication system.
2. GAN-based attacks. Due to the advent of GAN, researchers have also attempted to leverage GAN to craft adversarial examples.
    - AdvGAN. Xiao et al. [141] proposed AdvGAN to attack models by leveraging generative adversarial networks (GANs) [142], which can learn and approximate the distribution of original instances. After training the generator, AdvGAN can craft perturbations efficiently for any instance to accelerate AT as a defense potentially.
    - AdvFaces. Deb et al. [143] employed GAN to craft adversarial faces to evade face recognition software. AdvFaces comprises a generator, a discriminator, and a face matcher; therein, the fully convolution network was employed as a patch-based discriminator [144].

3.2. Physical world attacks

Physical world attacks refer to attacks against physical objects or systems by exploiting vulnerabilities in DL models, which have received increasing attention as DL models are increasingly used in safety-critical applications. These attacks involve manipulating the input data fed into a DL model, causing it to misclassify or produce unexpected results. For example, adversaries could use a sticker or other physical modification to trick a self-driving vehicle's DL model into misidentifying a stop sign as a yield sign [145], which leads to severe consequences if the vehicle continues through the intersection without stopping. Conducting physical world attacks generally consist of four steps [146]: crafting perturbations in the digital space, manufacturing in the physical space, capturing with sensors in the physical space, and attacking in the physical space. Physical attacks are carried out through a variety of adversarial media. We list physical world attacks based on different adversarial media in brief. For a more detailed introduction, readers can refer to a comprehensive survey of physical world attacks in [147].

1. Stickers and patches. This type of attack is the most common physical world attack. Adversaries attach stickers or patches to the surface of the attacked object, where the stickers tend to be patterned and irregularly shaped, while patches are usually regular in shape. Through adversarial stickers, researchers proposed a variety of approaches, including CAMOU [148], adversarial camera sticker [149], Advhat [150], ER attack [151], translucent patch [152], dual attention suppression [153], full-coverage camouflage attack [154], differentiable transformation network [155], and adversarial color film [156]. Through adversarial patches, researchers designed adversarial patch [157], robust physical perturbations [145], perceptual-sensitive GAN [158], adversarial YOLO [159], AdvPattern [160], AdvArcFace [161], adversarial automatic check-out [162], and physical attack on monocular depth estimation with optimal adversarial patches [163].
2. Clothing, eyeglasses, and makeup. This method is used on people, and both adversarial clothing [164,165] and eyeglasses [166] are wearable. Adversarial clothing is used to trick person detectors into hiding under surveillance systems. Adversarial eyeglasses and makeup [167] can fool face recognition systems. However, adversarial eyeglasses and makeup are often exaggerated and not very concealed.
3. Image and 3D-printed object. The adversarial images in the physical world are distributed over the whole image to mislead the DL model. Kurakin et al. [113] printed out adversarial examples generated by FGSM and BIM and then took these pictures using cell phones, regardless of lighting conditions or shooting angle. The pictures produced by this method are still adversarial examples. Using 3D printing technology can also manufacture adversarial examples in the physical world. Athalye et al. [168] successfully printed 3D adversarial turtles. How to hide the perturbation while ensuring the attack effectiveness effectively is challenging.
4. Light. Light from a laser pointer [169], projector [170,171], or light bulbs [172] can also influence the recognition of objects. For example, Li et al. [173] conducted optical adversarial attacks on 3D face recognition system through an projector. However, these methods would be degraded in an environment with intense light. Moreover, the above-mentioned methods are all optical interference to the objects with artificial equipment, which are not very stealthy. Recently, Zhong et al. [174] have used shadows to attack traffic signs, where shadows are generated by cardboard. This attack looks more natural but requires a strong light source to be effective, and this method does not currently support targeted attacks.
5. Camera. This method [175] attacks by modifying the rolling shutter effect or image signal processing of a camera. The camera-based attack is very stealthy because it occurs on the camera, not on the object being attacked.

4. Defense techniques

Since the DL model was found to be sensitive to deliberately designed adversarial samples [19], scholars have set off a frenzy of research on adversarial attacks and defenses. During this decade, thousands of articles designed adversarial attack methods from different perspectives, and correspondingly, many defense methods were proposed to deal with adversarial attacks. Adversarial DL enters arms race mode. This section will review the existing reviews in defense techniques first. Then, we provide a comprehensive guide on defense approaches based on summarizing the classification of defense techniques in the current reviews. Limited by the length of the article and the emergence of new defense methods, this survey does not list all the defenses.

Akhtar and Mian [24] divided defense strategies into modifying data, modifying models, and using additional tools. Subsequently, Qiu et al. [108] followed this category to review defense techniques. According to the phase in which the defense occurs, Yuan et al. [26] summarized the defense techniques into proactive countermeasures, reactive countermeasures, and ensemble defenses. Different from previous categories, Xu et al. [110] presented the defense strategies on

Table 4
A summary of the categories and characteristics of defense methods introduced in Section 4.

Defense strategies	Characteristics	Defense techniques	References
Detect adversarial examples	Detect adversarial samples using the differences in characteristics or results between adversarial and clean samples	Train a classifier	[178–182]
		Train an autoencoder	[183]
		Scalar quantization and smoothing spatial filters	[184]
		Construct robust algorithms	[185]
		Feature squeezing	[186]
		Perturbation rectifying network	[187]
		Perceptual image hashing scheme	[188]
Weaken adversarial examples	Eliminate the impact of adversarial examples using compression, dimensionality reduction, adding noise or filtering	Foveation mechanism	[189]
		Data transformation	[190]
		JPG compression	[191–193]
		Principal component analysis	[194]
		GAN	[195]
		Bounded ReLU	[196]
		Additive noise	[197]
		Non-invertible data transformation	[198,199]
		Statistical filtering	[200]
		Gaussian data augmentation	[196]
Strengthen models	Improve the inherent robustness of the model using various approaches	Adversarial training	[40,201–207]
		Gradient regularization	[208–210]
		Gradient masking	[211]
		Defensive distillation	[212–214]
		Robust feature extraction	[215–219]
		Deep contractive networks	[41]
		Manifold regularized networks	[220]
		Ensemble methods	[221]
		Overcomplete output layer	[222]
		Layer-wise regularization	[223]
		Add noise to the logit outputs of networks	[224]
Search for robust architectures	Explore which architectures are relative robust	Search for hyperparameters	[225–227]
		Differentiable NAS	[228–231]
		Evolutionary algorithm	[232–237]
		Random sampling	[238,239]
		Anti-bandit strategy	[240]
		Deep pursuit algorithm	[241]

gradient masking, robust optimization, and adversarial example detection. Similar to [110], Wang et al. [176] divided defense strategies into four categories, reducing or eliminating the adversarial attack, enhancing the robustness of the network, ensemble learning, and detecting only. Instead of classifying various defense techniques, Chakraborty et al. [177] reviewed the popular countermeasures, including adversarial training, gradient hiding, defensive distillation, feature squeezing, blocking the transferability, the defense-GAN, MagNet, using HGD, and using basis function transformations. In some reviews, robust verification is also considered a defensive technique. In this survey, we consider network verification one of the robustness evaluation criteria for networks, which has been presented in Section 2.3.2.

Based on [176], we improve the taxonomy and divide defense techniques into four categories: (1) detection of adversarial examples; (2) weakening the adversarial examples; (3) strengthening the model; (4) search for robust architectures. The categories and characteristics of defense methods are summarized in Table 4. This taxonomy comprehensively encompasses a variety of defense methods, and we hope that subsequent research on defense techniques will fall into the four categories of approaches we present.

4.1. Detecting adversarial examples

The most straightforward way to defend against attacks is to detect them. Since the characteristics of adversarial samples and clean samples are different, adversarial examples can be distinguished by classification models. Metzen et al. [178] proposed an adversary detection network by training a network on binary classification tasks to distinguish between original images and adversarial examples. In addition, Lu et al. [179] introduced SafetyNet, which detects adversarial examples using an RBF-SVM on binary or quaternary codes. Both adversary detection network and SafetyNet work at the final output layer and can detect adversarial examples generated by FGSM [40], BIM [113], and Deepfool [118]. Working on intermediate layers, Li and Li [180] used statistics on convolutional layer outputs and designed a cascade classifier to identify adversarial examples. Additionally, Grosse et al. [181] and Hosseini et al. [182] added an additional class to the targeted classification model to classify adversarial examples. Therein, Hosseini et al. [182] used this approach to detect black-box attacks.

Besides, a variety of approaches have been leveraged to detect adversarial examples. For example, MagNet [183] trains an autoencoder as the detector and uses reconstruction error to estimate the distance between the manifold of benign examples and the input. However, MagNet could be fooled if perturbations are larger [242]. Moreover, scalar quantization and smoothing spatial filters [184] can also be employed to detect adversarial examples. Feinman et al. [243] combined kernel density estimates in the subspace of the last hidden layer, and Bayesian neural network uncertainty estimates to detect adversarial perturbations. By applying persistent homology to the induced graphs, Gebhart and Schrater [185] constructed three robust algorithms to detect adversarial inputs. However, they only conducted experiments on MNIST [244] instead of larger datasets like CIFAR-10.

Since adversarial interference will be reduced after feature squeezing, Xu et al. [186] reduced the dimensionality of the image by reducing the color bit depth of each pixel and performing spatial smoothing over the image. If the dimensionality reduction image classification result is inconsistent with the previous one, the image is considered an adversarial example. Akhtar et al. [187] proposed a method to defend against universal adversarial attacks [121]. A perturbation rectifying network (PRN), learned from original and adversarial examples, is used as pre-input layers to a targeted model. A detector is trained on the discrete cosine transform of the input–output difference of the PRN. If adversarial perturbations are detected, the output of PRN is used to classify the image. To detect query-based adversarial attacks, Choi et al. [188] proposed PIHA, which generates a hash sequence using a perceptual image hashing scheme and compares the hash sequence with previous queries.

4.2. Weakening adversarial examples

In addition to detecting adversarial examples, many studies have also focused on devising strategies to reduce the effectiveness of adversarial attacks. Luo et al. [189] presented that the impact of adversarial attacks using L-BFGS [19] and FGSM [40] could be mitigated by the foveation mechanism in different image regions. They observed that the DNN model is robust to scale changes and transformations of the original image. This is a property that cannot be generalized to adversarial models. Therefore, the foveation serves as a solution to reduce the impact of adversarial attacks. However, the effectiveness of the foveation against more powerful attacks is yet to be demonstrated.

After that, Wang et al. [190] used a data transformation module to transform the input data to a new representation, which increases the complexity of a DNN model, augments its resistance to adversarial examples, and blocks the backward flows of gradients.

To weaken the attacks, Dziugaite et al. [191], Guo et al. [192], and Das et al. [193] employed JPG compression on the adversarial examples. Nevertheless, Shin and Song [245] demonstrated adversarial examples can survive JPEG compression. More significant compressions also lead to poor results on clean images, while minor compressions are often not able to remove the adversarial perturbations. Additionally, Bhagoji et al. [194] employed principal component analysis to compress data. However, Xu et al. [246] noted that this compression also corrupts the spatial structure of the image, often affecting performance.

Besides, a variety of techniques are proposed to weaken adversarial images. For example, Shen et al. [195] proposed adversarial perturbation elimination with GAN to rectify perturbed images. Zantedeschi et al. [196] adopted bounded ReLU [247] to reduce the effectiveness of adversarial patterns. Jin et al. [197] introduced a feedforward CNN leveraging additive noise to weaken adversarial examples. Sun et al. [200] proposed HyperNetworks that use statistical filtering to robustify the network. By integrating an input transformation to the DNN model, Wang et al. [198,199] leveraged non-invertible data transformation to conduct dimensionality reduction to protect DNNs from adversarial attacks. Gaussian data augmentation [196] also slightly helps improve the model robustness against adversarial attacks.

4.3. Strengthening models

Detecting adversarial examples and weakening adversarial examples do not make the model itself more robust. Strengthening the model is the most important and effective method of robust DL.

1. Adversarial training. Adversarial training (AT) augments the original training data with adversarial examples to build a more robust model. It is one of the most widely used defense techniques, which helps regularize the network to reduce overfitting [40,201] and makes the model more resistant to adversarial attacks. Na et al. [202] used an already defended network for adversarial image generation by BIM. Since generating adversarial samples is inherently time-consuming, AT takes more time than standard training, which makes it intractable to use the method on large-scale datasets, which is an urgent problem for AT defense methods. Researchers attempted different ways to improve AT to reduce computational costs. For example, Shafahi et al. [203] improved the efficiency of AT by updating the model parameters and image perturbations simultaneously using one backward pass. Since obtaining the gradients that produce adversarial samples is computationally expensive, researchers also tempted to approximate these gradients and then train the model. For instance, Miyato et al. [204] proposed a virtual AT method that can be used for supervised and semi-supervised learning tasks. In addition to virtual AT, stability learning methods [205], leveraging the divergence-based distributional robustness of the model against the virtual adversarial direction to represent local distributional smoothness, can also contribute to strengthening models.
   Lee et al. [206] were the first to combine the framework of GAN with AT. They employed the generator of GAN to create adversarial examples, then used a classifier to train the original and generated examples. This process does not stop until the generator cannot generate adversarial examples that are classifier misclassified. They also found this method is effective in regularizing neural networks. In addition to training the network with adversarial examples, Li et al. [207] proposed squeeze training, which was trained together with adversarial examples and collaborative examples, where the collaborative examples are the examples with a lower predicted loss in the ϵ bounded neighborhood of the clean example.
2. Gradient regularization/masking. In addition to enhancing the model robustness through adversarial training, several other studies have concentrated on gradient regularization or masking. Lyu et al. [208] used gradient regularization methods that penalize the gradient of loss functions with respect to the inputs. Similarly, Shaham et al. [209] improved the local stability of models by minimizing the loss of the model over worst-case adversarial examples at each parameter update. Besides, DeepCloak [211] inserts a mask before the fully connected layer to remove unnecessary information from the network to reduce the interference of adversarial examples on the model. Moreover, Ross and Doshi-Velez [210] leveraged input gradient regularization [248] to make small adversarial perturbations become difficult to change the output of the trained model significantly.
3. Defensive distillation. Another way to efficiently defend against adversarial attacks is defensive distillation. Papernot et al. [212] exploited defensive distillation [249] that transfers knowledge of a complex network to a smaller network. The defensive distillation technique starts by training a teacher network from a given temperature and obtaining the soft labels of the training data. Then the soft labels are used to train a distilled network. The distillation model is used to classify adversarial samples to improve the robustness of the model. Further empirical evidence is also provided in [213]. Besides, Papernot and McDaniel [214] also enhanced defensive distillation by addressing the numerical instabilities encountered in [212].
4. Robust feature extraction. Based on the invariance acquired through domain knowledge or provided by real-world constraints, Chandrasekaran et al. [215] designed a new hierarchical classification method. Freitas et al. [216] proposed a feature alignment method to build robust DL models by mining deep semantic features from images and comparing them with the expected features of classification. Inspired by the way humans recognize objects, Li et al. [217] and Sitawarin et al. [218] employed part-based models to segment each component of the image, then the predictions can be made based on the parts. Ding et al. [219] designed a shallow binary feature module (SBFM) to extract the contour features of an image. SBFM consists of a sober layer and a threshold layer, and the edge features learned by this module are transmitted to the fully connected layer of the backbone network, and the input images are classified together with the features learned by the backbone network.
5. Other approaches. Except for the above methods, some methods are difficult to fall into the categories described above. Inspired by the contractive autoencoder [250], Gu and Rigazio [41] proposed deep contractive networks to regularize training by adding a penalty for partial derivatives of each layer in a standard backpropagation framework. In this way, the change of input data will not cause significant changes to the output of each layer, thereby ensuring the stability of the model classification prediction results. Later, Lee et al. [220] introduced manifold regularized networks that incorporate a manifold loss term with

the objective of minimizing the difference between multi-layer embedding results of original and adversarial samples. In 2017, Strauss et al. [221] investigated some ensemble methods to defend against perturbations, which make predictions by letting each classifier vote for a label. In the same year, Kadran and Stanley [222] introduced a competitive overcomplete output layer to induce robustness against adversarial attacks. Concurrently, Parseval networks Cisse et al. [223] employed a layer-wise regularization by controlling the global Lipschitz constant of the network. Subsequently, Nguyen et al. [224] introduced a masking-based defense by adding noise to the logit outputs of networks against low distortion attacks such as the C&W attack.

4.4. Search for robust architectures

The above methods improve the robustness of DL models by strengthening the model or adding additional models. In recent years, researchers have explored robust architectures through search. Robust architecture search can be divided into model hyperparameter search and topology search.

1. Search for hyperparameters for robustness.
   Considering adversarial robustness as one objective in a dual-objective optimization problem, Liu and Jin [225] first employed the elitist non-dominated sorting genetic algorithm (NSGA-II) [251] to optimize hyperparameters, including batch size, learning rate and momentum, against FGSM [40]. Alparslan and Kim [226] investigated the relationship between robustness and model size. Through a series of experiments, Huang et al. [227] leveraged grid search to explore the influence of depth and width on the robustness of the adversarially trained DNNs.
2. Search for robust model topologies.
   Neural architecture search aims to use a specific search method to search for a network architecture that satisfies specific objectives in a given search space. The objectives are usually the accuracy of the network on the specific task, the latency of the model, and so on. Robust architecture search can also be formulated as an optimization problem. We will then review the literature on robust NAS based on the search strategies used by the researchers.

   - Differentiable NAS. Differentiable NAS methods are commonly used in robust architecture search due to their high efficiency. For the adversarial medical image segmentation task, Dong et al. [228] employed differentiable NAS to search for the architecture of a discriminator. Based on certified lower bound and Jacobian norm bound, Hosseini et al. [229] proposed two robustness metrics and introduced DSRNA to maximize the proposed robustness metrics. Besides, Mok et al. [230] introduced AdvRush, adopting DARTS [252] as a backbone, to search for architectures with a smooth input loss landscape. For AI-enabled Internet-of-Things systems, Wang et al. [231] presented a multiobjective gradient optimization method and designed a new search space for robust NAS recently.
   - Evolutionary algorithm (EA). In a broader search space, robust architecture search (RAS) [232] leveraged an EA to search for architectures robust to transferable attacks. Focused on networks with small model sizes, Xie et al. [233] introduced TAM-NAS by employing one-shot NAS and NSGA-II [251] to search for the architectures with high clean accuracy, high adversarial accuracy, and tiny mode size. To search for robust architectures against various attacks, Liu and Jin [234] introduced MORAS to search for architectures that are less sensitive to five widely used adversarial attacks. Moreover, Liu et al. [235] further proposed MORAS-SH to use a surrogate model as an auxiliary objective to assist robust network architecture search in improving search efficiency. Employing the multiple-gradient descent algorithm, Yue et al. [236] presented E2RNAS to optimize the performance, the robustness, and the resource constraint, simultaneously. Aim at searching for robust architectures at targeted capacities, Ning et al. [237] proposed multi-shot NAS, employing a tournament-based evolutionary search strategy [253], architectures with high accuracies on clean and adversarial examples were obtained.
   - Random sampling. Based on AT and one-shot NAS, Guo et al. [238] introduced RobNet, which randomly samples the architectures and fine-tunes the sampled networks with AT for several epochs, reporting that densely connected structures benefit the model robustness. Besides, Devaguptapu et al. [239] explored robust architecture using random sampling on DARTS search space. However, no defense approach was used, so the accuracies under adversarial attacks were poor.
   - Other strategies. Except for the above search strategies, Chen et al. [240] presented ABanditNAS by designing a search space composed of denoising blocks, weight-free operations, Gabor filters and convolutions, and employing an anti-bandit strategy to search multiple cells based on the designed search space. In addition, Cazenavette et al. [241] proposed a deep pursuit algorithm that formulates the robust NAS as a global sparse coding problem, which jointly computes all network activations.

Table 5
A list of libraries for general robustness.

Benchmark	Link
ImageNet-A [254]	https://github.com/hendrycks/natural-adv-examples
ImageNet-O [254]	https://github.com/hendrycks/natural-adv-examples
ImageNet-C [255]	https://github.com/hendrycks/robustness
ImageNet-R [256]	https://github.com/hendrycks/imagenet-r
ImageNet-E [257]	https://github.com/alibaba/easyrobust

5. Open problems

Since DL models were found to be vulnerable to deliberately designed adversarial examples, rapidly increasing work has focused on robust DL in the past decade. Although the defense methods against attacks emerge endlessly, adversarial attacks are also diverse. Therefore, DL is still difficult to apply to crucial security-related fields. There remain many open questions about robust DL in CV.

1. Are white-box adversarial attacks easy to achieve in reality?
   Adversarial attacks have attracted considerable attention recently, and most attack methods achieve high success rates. Based on current digital research, extensive research focuses on designing white-box attacks and the corresponding defenses. White-box attacks assume adversaries know much knowledge, such as the gradient of the attacked network. Nevertheless, in practical scenarios, the parameters of the model may not be accessible to attackers. In addition, some of the more intense attacks, such as C&W and PGD-20, are time-consuming. It is easily detected if the adversary takes too long to attack. Based on the above two reasons, in real-world scenarios, how to design attacks that can be achieved in reality still needs to be studied. It is worth investigating and placing greater importance on black-box and physical world attacks in future work.
2. Are there general criteria for evaluating the robustness of DL models?
   When DNNs are used in safety-critical environments, especially against emerging adversary attacks, the robustness of DNNs

Table 6
List of libraries for adversarial robustness.

Toolbox	Link
AdvBox [258]	https://github.com/advboxes/AdvBox
Advertorch [259]	https://github.com/BorealisAI/advertorch
AISafety [46]	https://openi.pcl.ac.cn/OpenI/AISafety
ARES-Bench [260]	https://ml.cs.tsinghua.edu.cn/ares-bench/#/leaderboard
ART (Adversarial Robustness Toolbox) [261]	https://adversarial-robustness-toolbox.readthedocs.io/
CleverHans [262]	https://github.com/cleverhans-lab/cleverhans#setting-up-cleverhans
DEEPSEC [263]	https://github.com/ryderling/DEEPSEC
Foolbox [264]	https://foolbox.readthedocs.io/en/stable/#
RobustART [265]	http://robust.art/
RobustBench [266]	https://robustbench.github.io/

requires to be evaluated to help us understand how reliable the model's predictions are and how much we can rely on DL in practical scenarios. This study analyzes current robustness evaluation metrics. However, the formal standardization of robustness assessment methods for DL is still at the initial stage. For empirical solutions, certified robustness provides a formal alternative to obtain powerful DL models, but these certified solutions come at the cost of high computational complexity. Finding new low-complexity certified robustness more suitable for large models is an ongoing area of research. Therefore, it is vital to establish broadly applicable and well-defined robustness evaluation standards.

3. Are there benchmark platforms that can accelerate the development of robust DL?
For general robustness, Hendrycks et al. proposed a series of Benchmarks, such as ImageNet-C [255], ImageNet-R [256] and ImageNet-E [257] to test the robustness or generalization ability of the model by conducting common corruptions to the image or changing the background. The specific methods and links are shown in Table 5. For adversarial robustness, most attack and defense methods do not publicly describe the code of the technique, making it difficult for other researchers to efficiently and accurately replicate the solutions and deploy appropriate attacks and countermeasures. Since an increasing number of DL methods and robustness improvement techniques have been proposed, a comprehensive benchmark of model robustness is vital to understand their effectiveness and keep up with the state-of-the-art. Several libraries are available online to help researchers conduct attacks or defenses, including AdvBox [258], ART (Adversarial Robustness Toolbox) [261], Advertorch [259], ARES-Bench [260], CleverHans [262], DEEPSEC [263], Foolbox [264], RobustART [265] and RobustBench [266]. The links of the toolboxes are listed in Table 6. Besides, robust network architecture search is still in its infancy. The reason is that the NAS is time-consuming, doubling its time if the network is trained adversarially. And when testing network performance, attacking the network also costs a lot. This has deterred many researchers. If there is a Benchmark like NASBench-101 and NASBench-201, it will significantly promote the development and research in robust DL. Jung et al. [267] first provided a dataset based on NASBench-201 search space against four commonly used adversarial attacks and image dataset corruptions. This dataset is crucial to streamlining the research on robust NAS. The robustness of the architectures of the broader search space remains to be explored, such as the vision transformer [20].
4. Is there a trade-off between model robustness and generalization, interpretability, and privacy?
Generalization ability refers to the performance of the model on unseen data and is the primary indicator for evaluating network performance. Interpretability, privacy, and robustness are all essential components of trustworthy AI. Improving generalization performance, interpretability, and privacy are all conducive to broader applications of AI. However, improving them all at the same time is quite challenging. Researchers are exploring the relationship between robustness and them [268–270]. Does improving robustness help generalization performance, interpretability, and privacy, or are they conflicting goals? Scholars have yet to reach a conclusion. The research on the relationship between robustness and generalization performance, interpretability, and privacy needs to be deeper, and there is still a long way to be explored.

6. Conclusion

With the continuous deepening of DL research and the ever-increasing applications of DL technology in practical scenarios, the robustness of DL models has become a promising research field, calling the attention of numerous researchers in academia and industry. Scholars have a wide range of interests and in-depth research and have achieved excellent results. However, research on the robustness of DL models is still in its infancy, and many critical scientific issues remain to be resolved. To determine the current research status of the robustness of DL models, clarify the advantages and disadvantages of existing research results, and clarify future research directions, this paper systematically studies the robustness of DL models and reviews quantities of highly influential pieces of literature. We shed light on robust DL in CV to provide a comprehensive overview of this research field. Specifically, we focused on defining robustness, measuring robustness, describing the adversarial attacks and defenses, and summarizing the open problems that remain to be solved. We hope our work can be a reference to help researchers get a systematical and comprehensive understanding of robust DL in CV, thus providing more insights for their studies.

CRediT authorship contribution statement

Jia Liu: Data curation, Formal analysis, Investigation, Writing – original draft. **Yaochu Jin:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

Y. Jin is funded by an Alexander von Humboldt Professorship for Artificial Intelligence endowed by the German Federal Ministry of Education and Research.

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[2] Diego Marcheggiani, Ivan Titov, Encoding sentences with graph convolutional networks for semantic role labeling, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1506–1515.

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, Dong Yu, Convolutional neural networks for speech recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (10) (2014) 1533–1545.

[4] Matthew D. Zeiler, Rob Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[5] Min Lin, Qiang Chen, Shuicheng Yan, Network in network, 2013, arXiv preprint arXiv:1312.4400.

[6] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: Yoshua Bengio, Yann LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 4700–4708.

[10] Sara Sabour, Nicholas Frosst, Geoffrey E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, Jian Sun, Object detection networks on convolutional feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 39 (7) (2017) 1476–1481.

[13] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[14] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, Tian Xia, Multi-view 3d object detection network for autonomous driving, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[15] Peng Li, Jiabin Zhang, Zheng Zhu, Yanwei Li, Lu Jiang, Guan Huang, State-aware re-identification feature for multi-target multi-camera tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.

[17] Charlotte Middlehurst, China unveils world's first facial recognition ATM, Telegraph 1 (2015).

[18] Andrew Bud, Facing the future: The impact of apple faceid, Biom. Technol. Today 2018 (1) (2018) 5–7.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus, Intriguing properties of neural networks, in: Yoshua Bengio, Yann LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[21] Kaleel Mahmood, Rigel Mahmood, Marten Van Dijk, On the robustness of vision transformers to adversarial examples, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7838–7847.

[22] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, Carl Vondrick, Understanding zero-shot adversarial robustness for large-scale models, in: The Eleventh International Conference on Learning Representations, 2023.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[24] Naveed Akhtar, Ajmal Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, Ieee Access 6 (2018) 14410–14430.

[25] Mesut Ozdag, Adversarial attacks and defenses against deep neural networks: A survey, Procedia Comput. Sci. 140 (2018) 152–161.

[26] Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Trans. Neural Netw. Learn. Syst. 30 (9) (2019) 2805–2824.

[27] Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, Adversarial attacks and defenses in deep learning, Engineering 6 (3) (2020) 346–360.

[28] Samuel Henrique Silva, Peyman Najafirad, Opportunities and challenges in deep learning adversarial robustness: A survey, 2020, arXiv preprint arXiv:2007.00753.

[29] Muhammad Imran Tariq, Nisar Ahmed Memon, Shakeel Ahmed, Shahzadi Tayyaba, Muhammad Tahir Mushtaq, Natash Ali Mian, Muhammad Imran, Muhammad W Ashraf, A review of deep learning security and privacy defensive techniques, Mob. Inf. Syst. 2020 (2020).

[30] Jia Ding, Zhiwu Xu, Adversarial attacks on deep learning models of computer vision: A survey, in: Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part III 20, Springer, 2020, pp. 396–408.

[31] Teng Long, Qi Gao, Lili Xu, Zhangbing Zhou, A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions, Comput. Secur. 121 (2022) 102847.

[32] Jinyin Chen, Yan Zhang, Xueke Wang, Hongbin Cai, Jue Wang, Shouling Ji, A survey of attack, defense and related security analysis for deep reinforcement learning, Acta Automat. Sinica 48 (AAS-CN-2020-0166) (2022) 21.

[33] Sara Sabour, Yanshuai Cao, Fartash Faghri, David J Fleet, Adversarial manipulation of deep representations, in: ICLR (Poster), 2016.

[34] Andras Rozsa, Ethan M. Rudd, Terrance E. Boult, Adversarial diversity and hard positive generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 25–32.

[35] Daniel Zügner, Amir Akbarnejad, Stephan Günnemann, Adversarial attacks on neural networks for graph data, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2847–2856.

[36] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 Acm Sigsac Conference on Computer and Communications Security, 2016, pp. 1528–1540.

[37] Yaochu Jin, Bernhard Sendhoff, Trade-off between performance and robustness: an evolutionary multiobjective approach, in: International Conference on Evolutionary Multi-Criterion Optimization, Springer, 2003, pp. 237–251.

[38] Nathan Drenkow, Numair Sani, Ilya Shpitser, Mathias Unberath, Robustness in deep learning for computer vision: Mind the gap?, 2021, http://dx.doi.org/10.48550/ARXIV.2112.00639.

[39] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, Antonio Criminisi, Measuring neural net robustness with constraints, in: Advances in Neural Information Processing Systems, Vol. 29, 2016, pp. 2613–2621.

[40] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.

[41] Shixiang Gu, Luca Rigazio, Towards deep neural network architectures robust to adversarial examples, 2014, arXiv preprint arXiv:1412.5068.

[42] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, Mykel J Kochenderfer, Reluplex: An efficient SMT solver for verifying deep neural networks, in: International Conference on Computer Aided Verification, Springer, 2017, pp. 97–117.

[43] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, Mykel J Kochenderfer, Towards proving the adversarial robustness of deep neuralnetworks, 2017, arXiv preprint arXiv:1709.02126.

[44] Ravi Mangal, Aditya V. Nori, Alessandro Orso, Robustness of neural networks: a probabilistic and practical approach, in: Anita Sarma, Leonardo Murta (Eds.), Proceedings of the 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE (NIER) 2019, Montreal, QC, Canada, May 29-31, 2019, IEEE / ACM, 2019, pp. 93–96.

[45] Natan Levy, Guy Katz, Roma: a method for neural network robustness measurement and assessment, 2021, arXiv preprint arXiv:2110.11088.

[46] Jun Guo, Wei Bao, Jiakai Wang, Yuqing Ma, Xinghai Gao, Gang Xiao, Aishan Liu, Jian Dong, Xianglong Liu, Wenjun Wu, A comprehensive evaluation framework for deep model robustness, Pattern Recognit. 137 (2023) 109308.

[47] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, Jun Zhu, Benchmarking adversarial robustness on image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 321–331.

[48] Chih-Ling Chang, Jui-Lung Hung, Chin-Wei Tien, Chia-Wei Tien, Sy-Yen Kuo, Evaluating robustness of AI models against adversarial attacks, in: Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence, 2020, pp. 47–54.

[49] Changliu Liu, Tomer Arnon, Christopher Lazarus, Christopher Strong, Clark Barrett, Mykel J. Kochenderfer, Algorithms for verifying deep neural networks, Found. Trends® Optim. 4 (3–4) (2021) 244–404.

[50] S Ji, T Du, S Deng, P Cheng, J Shi, M Yang, B Li, Robustness certification research on deep learning models: A survey, Chin. J. Comput. 45 (2022) 190–206.

[51] Linyi Li, Tao Xie, Bo Li, SoK: Certified robustness for deep neural networks, in: 44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023, IEEE, 2023.

[52] Chih-Hong Cheng, Georg Nührenberg, Harald Ruess, Maximum resilience of artificial neural networks, in: Deepak D'Souza, K. Narayan Kumar (Eds.), Automated Technology for Verification and Analysis, Springer International Publishing, Cham, 2017, pp. 251–268.

[53] Ignacio E. Grossmann, Review of nonlinear mixed-integer and disjunctive programming techniques, Optim. Eng. 3 (2002) 227–252.

[54] Patrick Cousot, Radhia Cousot, Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints, in: Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages, POPL '77, Association for Computing Machinery, New York, NY, USA, 1977, pp. 238–252.

[55] Eric Wong, J. Zico Kolter, Provable defenses against adversarial examples via the convex outer adversarial polytope, in: Jennifer G. Dy, Andreas Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, in: Proceedings of Machine Learning Research, 80, PMLR, 2018, pp. 5283–5292.

[56] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A. Mann, Pushmeet Kohli, A dual approach to scalable verification of deep networks, in: Amir Globerson, Ricardo Silva (Eds.), Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018, AUAI Press, 2018, pp. 550–559.

[57] Aditi Raghunathan, Jacob Steinhardt, Percy Liang, Certified defenses against adversarial examples, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.

[58] Aditi Raghunathan, Jacob Steinhardt, Percy Liang, Semidefinite relaxations for certifying robustness to adversarial examples, in: Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, MontrÉAl, Canada, 2018, pp. 10900–10910.

[59] Mahyar Fazlyab, Manfred Morari, George J. Pappas, Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming, IEEE Trans. Automat. Control 67 (2019) 1–15.

[60] Matt Jordan, Justin Lewis, Alexandros G. Dimakis, Provable certificates for adversarial examples: Fitting a ball in the union of polytopes, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.

[61] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, Pengchuan Zhang, A convex relaxation barrier to tight robustness verification of neural networks, Adv. Neural Inf. Process. Syst. 32 (2019).

[62] Luca Pulina, Armando Tacchella, An abstraction-refinement approach to verification of artificial neural networks, in: Proceedings of the 22nd International Conference on Computer Aided Verification, CAV '10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 243–257.

[63] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, Martin Vechev, AI2: Safety and robustness certification of neural networks with abstract interpretation, in: 2018 IEEE Symposium on Security and Privacy, (SP), 2018, pp. 3–18.

[64] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, Martin T. Vechev, Fast and effective robustness certification, in: Neural Information Processing Systems, 2018.

[65] Matthew Mirman, Timon Gehr, Martin T. Vechev, Differentiable abstract interpretation for provably robust neural networks, in: International Conference on Machine Learning, 2018.

[66] Gagandeep Singh, Timon Gehr, Markus Püschel, Martin Vechev, An abstract domain for certifying neural networks, Proc. ACM Program. Lang. 3 (POPL) (2019).

[67] Gagandeep Singh, Timon Gehr, Markus Püschel, Martin T. Vechev, Boosting robustness certification of neural networks, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, la, USA, May 6-9, 2019, OpenReview.net, 2019.

[68] Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, Martin Vechev, Beyond the single neuron convex barrier for neural network certification, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 32, Curran Associates, Inc., 2019.

[69] Matthias Hein, Maksym Andriushchenko, Formal guarantees on the robustness of a classifier against adversarial manipulation, in: Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S.V.N. Vishwanathan, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 2266–2276.

[70] Wenjie Ruan, Xiaowei Huang, Marta Kwiatkowska, Reachability analysis of deep neural networks with provable guarantees, in: Jérôme Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 2651–2659.

[71] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane S. Boning, Inderjit S. Dhillon, Towards fast computation of certified robustness for ReLU networks, in: Jennifer G. Dy, Andreas Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, StockholmsmäSsan, Stockholm, Sweden, July 10-15, 2018, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 5273–5282.

[72] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, Luca Daniel, Evaluating the robustness of neural networks: An extreme value theory approach, 2018, arXiv preprint arXiv:1801.10578.

[73] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Aurelie Lozano, Cho-Jui Hsieh, Luca Daniel, On extensions of CLEVER: A neural network robustness evaluation algorithm, in: IEEE Global Conference on Signal and Information Processing, (GlobalSIP), 2018.

[74] Jean Bernard Lasserre, An Introduction To Polynomial and Semi-Algebraic Optimization, in: Cambridge Texts in Applied Mathematics, Cambridge University Press, 2015.

[75] Fabian Latorre Gómez, Paul Rolland, Volkan Cevher, Lipschitz constant estimation of neural networks via sparse polynomial optimization, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[76] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Suman Jana, Certified robustness to adversarial examples with differential privacy, in: 2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019, IEEE, 2019, pp. 656–672.

[77] Jeremy M. Cohen, Elan Rosenfeld, J. Zico Kolter, Certified adversarial robustness via randomized smoothing, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 1310–1320.

[78] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, Jamal Atif, Theoretical evidence for adversarial robustness through randomization, in: Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 11838–11848.

[79] Guang-He Lee, Yang Yuan, Shiyu Chang, Tommi S. Jaakkola, Tight certificates of adversarial robustness for randomly smoothed classifiers, in: Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 4911–4922.

[80] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan L. Yuille, Mitigating adversarial effects through randomization, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.

[81] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, Animashree Anandkumar, Stochastic activation pruning for robust adversarial defense, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.

[82] Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, Greg Yang, Provably robust deep learning via adversarially trained smoothed classifiers, in: Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 11289–11300.

[83] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, J. Zico Kolter, Chongli Qin, András György, Kai Xiao, Sven Gowal, Pushmeet Kohli, A framework for robustness certification of smoothed classifiers using F-divergences, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[84] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, J. Zico Kolter, Black-box smoothing: A provable defense for pretrained classifiers, CoRR (2020).

[85] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, Neil Zhenqiang Gong, Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[86] Binghui Wang, Xiaoyu Cao, Jinyuan Jia, Neil Zhenqiang Gong, On certifying robustness against backdoor attacks via randomized smoothing, CoRR (2020) arXiv:2002.11750, arXiv:2002.11750, URL https://arxiv.org/abs/2002.11750.

[87] Maurice Weber, Xiaojun Xu, Bojan Karlas, Ce Zhang, Bo Li, RAB: provable robustness against backdoor attacks, in: 2023 IEEE Symposium on Security and Privacy, (SP), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 640–657.

[88] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Sijia Liu, Pin-Yu Chen, Luca Daniel, Rethinking randomized smoothing for adversarial robustness, CoRR (2020).

[89] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, Suman Jana, Formal security analysis of neural networks using symbolic intervals, in: Proceedings of the 27th USENIX Conference on Security Symposium, SEC '18, USENIX Association, USA, 2018, pp. 1599–1614.

[90] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, Suman Jana, Efficient formal safety analysis of neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS '18, Curran Associates Inc., Red Hook, NY, USA, 2018, pp. 6369–6379.

[91] Teruo Sunaga, Theory of an interval algebra and its application to numerical analysis, Japan J. Ind. Appl. Math. 26 (2009) 125–143.

[92] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, Pushmeet Kohli, Scalable verified training for provably robust image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4842–4851.

[93] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelović, Brendan O'Donoghue, Jonathan Uesato, Pushmeet Kohli, Training verified learners with learned verifiers, ArXiv (2018) arXiv:1805.10265.

[94] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane S. Boning, Cho-Jui Hsieh, Towards stable and efficient training of verifiably robust neural networks, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[95] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, Luca Daniel, Efficient neural network robustness certification with general activation functions, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.

[96] Yuh-Shyang Wang, Tsui-Wei Weng, Luca Daniel, Verification of neural network control policy under persistent adversarial perturbation, CoRR (2019).

[97] Bao Wang, Zuoqiang Shi, Stanley J. Osher, ResNets ensemble via the feynman-kac formalism to improve natural and robust accuracies, in: Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 1655–1665.

[98] Steven Carr, Nils Jansen, Ufuk Topcu, Verifiable RNN-based policies for POMDPs under temporal logic constraints, in: Christian Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 4121–4127.

[99] Lily Weng, Pin-Yu Chen, Lam M. Nguyen, Mark S. Squillante, Akhilan Boopathy, Ivan V. Oseledets, Luca Daniel, PROVEN: verifying robustness of neural networks with a probabilistic approach, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6727–6736.

[100] Mahyar Fazlyab, Manfred Morari, George J. Pappas, Probabilistic verification and reachability analysis of neural networks via semidefinite programming, in: 2019 IEEE 58th Conference on Decision and Control, (CDC), 2019, pp. 2726–2731.

[101] S. Webb, T. Rainforth, Y. Teh, P. Mudigonda, A statistical approach to assessing neural network robustness, in: Seventh International Conference on Learning Representations, (ICLR 2019), International Conferences on Learning Representations, 2019.

[102] Divya Gopinath, Guy Katz, Corina S Păsăreanu, Clark Barrett, Deepsafe: A data-driven approach for assessing robustness of neural networks, in: International Symposium on Automated Technology for Verification and Analysis, Springer, 2018, pp. 3–19.

[103] Yue-Huan Wang, Ze-Nan Li, Jing-Wei Xu, Ping Yu, Taolue Chen, Xiao-Xing Ma, Predicted robustness as QoS for deep neural network models, J. Comput. Sci. Tech. 35 (5) (2020) 999–1015.

[104] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, Alexey Kurakin, On evaluating adversarial robustness, 2019, arXiv preprint arXiv:1902.06705.

[105] Alexandru Constantin Serban, Erik Poll, Joost Visser, Adversarial examples-a complete characterisation of the phenomenon, 2018, arXiv preprint arXiv:1810.01185.

[106] Alex Serban, Erik Poll, Joost Visser, Adversarial examples on object recognition: A comprehensive survey, ACM Comput. Surv. 53 (3) (2020) 1–38.

[107] Naveed Akhtar, Ajmal Mian, Navid Kardan, Mubarak Shah, Advances in adversarial attacks and defenses in computer vision: A survey, IEEE Access 9 (2021) 155161–155196.

[108] Shilin Qiu, Qihe Liu, Shijie Zhou, Chunjiang Wu, Review of artificial intelligence adversarial attack and defense technologies, Appl. Sci. 9 (5) (2019) 909.

[109] Yiyun Zhou, Meng Han, Liyuan Liu, Jing He, Xi Gao, The adversarial attacks threats on computer vision: A survey, in: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops, (MASSW), IEEE, 2019, pp. 25–30.

[110] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, Anil K Jain, Adversarial attacks and defenses in images, graphs and text: A review, Int. J. Autom. Comput. 17 (2020) 151–178.

[111] Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, Feng Li, A survey on adversarial attack in the age of artificial intelligence, Wirel. Commun. Mob. Comput. 2021 (2021) 1–22.

[112] Teng Long, Qi Gao, Lili Xu, Zhangbing Zhou, A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions, Comput. Secur. (2022) 102847.

[113] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Adversarial examples in the physical world, 2016, arXiv preprint arXiv:1607.02533.

[114] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

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

[116] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, John E Hopcroft, Nesterov accelerated gradient and scale invariance for adversarial attacks, 2019, arXiv preprint arXiv:1908.06281.

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

[118] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

[119] Nicholas Carlini, David Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy, (SP), IEEE, 2017, pp. 39–57.

[120] Chris Finlay, Aram-Alexandre Pooladian, Adam Oberman, The logbarrier adversarial attack: making effective use of decision boundary information, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4862–4870.

[121] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, Universal adversarial perturbations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1765–1773.

[122] Jorge Nocedal, Stephen J. Wright, Numerical Optimization, Springer, 1999.

[123] Yanpei Liu, Xinyun Chen, Chang Liu, Dawn Song, Delving into transferable adversarial examples and black-box attacks, in: International Conference on Learning Representations, 2017.

[124] Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Evading defenses to transferable adversarial examples by translation-invariant attacks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4312–4321.

[125] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, Alan L Yuille, Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.

[126] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge J. Belongie, Ser-Nam Lim, Enhancing adversarial example transferability with an intermediate level attack, ArXiv (2019) arXiv:1907.10823.

[127] Zhichao Huang, Tong Zhang, Black-box adversarial attack with transferable model-based embedding, in: International Conference on Learning Representations, 2020.

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

[129] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 15–26.

[130] Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin, Black-box adversarial attacks with limited queries and information, in: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, 2018.

[131] Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, Aaron van den Oord, Adversarial risk and the dangers of evaluating against weak attacks, in: Jennifer Dy, Andreas Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 5025–5034.

[132] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, Boqing Gong, NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 3866–3876.

[133] Wieland Brendel, Jonas Rauber, Matthias Bethge, Decision-Based Adversarial, Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, Adv. Reliab. Eval. Improv. Adversarial Robust. (2021) 77.

[134] Jiawei Su, Danilo Vasconcellos Vargas, Kouichi Sakurai, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput. 23 (5) (2019) 828–841.

[135] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, Jun Zhu, Efficient decision-based black-box adversarial attacks on face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7714–7722.

[136] Shuai Jia, Yibing Song, Chao Ma, Xiaokang Yang, IoU attack: Towards temporally coherent black-box adversarial attack for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6709–6718.

[137] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, Ilya Sutskever, Evolution strategies as a scalable alternative to reinforcement learning, 2017, arXiv preprint arXiv:1703.03864.

[138] James C. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE Trans. Autom. Control 37 (3) (1992) 332–341.

[139] Christian Igel, Thorsten Suttorp, Nikolaus Hansen, A computational efficient covariance matrix update and a (1+ 1)-CMA for evolution strategies, in: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, 2006, pp. 453–460.

[140] Hanrui Wang, Shuo Wang, Zhe Jin, Yandan Wang, Cunjian Chen, Massimo Tistarelli, Similarity-based gray-box adversarial attack against deep face recognition, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, (FG 2021), IEEE Press, 2021, pp. 1–8.

[141] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, Dawn Song, Generating adversarial examples with adversarial networks, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI '18, AAAI Press, 2018, pp. 3905–3911.

[142] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[143] Debayan Deb, Jianbang Zhang, Anil K. Jain, AdvFaces: Adversarial face synthesis, in: 2020 IEEE International Joint Conference on Biometrics, (IJCB), 2020, pp. 1–10.

[144] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2017, pp. 5967–5976.

[145] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, Robust physical-world attacks on deep learning visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[146] Hui Wei, Hao Tang, Xuemei Jia, Han-Bing Yu, Zhubo Li, Zhixiang Wang, Shin'ichi Satoh, Zheng Wang, Physical adversarial attack meets computer vision: A decade survey, ArXiv (2022) arXiv:2209.15179.

[147] Hui Wei, Hao Tang, Xuemei Jia, Hanxun Yu, Zhubo Li, Zhixiang Wang, Shin'ichi Satoh, Zheng Wang, Physical adversarial attack meets computer vision: A decade survey, 2022, arXiv preprint arXiv:2209.15179.

[148] Yang Zhang, Hassan Foroosh, Phiip David, Boqing Gong, CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild, in: International Conference on Learning Representations, 2018.

[149] Juncheng Li, Frank Schmidt, Zico Kolter, Adversarial camera stickers: A physical camera-based attack on deep learning systems, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 3896–3904.

[150] Stepan Komkov, Aleksandr Petiushko, AdvHat: Real-world adversarial attack on ArcFace face ID system, in: 2020 25th International Conference on Pattern Recognition, (ICPR), 2021, pp. 819–826.

[151] Tong Wu, Xuefei Ning, Wenshuo Li, Ranran Huang, Huazhong Yang, Yu Wang, Physical adversarial attack on vehicle detector in the carla simulator, ArXiv (2020) arXiv:2007.16118.

[152] Alon Zolfi, Moshe Kravchik, Yuval Elovici, Asaf Shabtai, The translucent patch: A physical and universal attack on object detectors, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 15227–15236.

[153] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, Xianglong Liu, Dual attention suppression attack: Generate adversarial camouflage in physical world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8565–8574.

[154] Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Xiaoya Zhang, Zhiqiang Gong, Wen Yao, Xiaoqian Chen, FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack, CoRR (2021) arXiv:2109.07193, arXiv:2109.07193, URL https://arxiv.org/abs/2109.07193.

[155] Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, Howon Kim, DTA: Physical camouflage attacks using differentiable transformation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15305–15314.

[156] Chengyin Hu, Weiwen Shi, Adversarial color film: Effective physical-world attack to DNNs, 2023.

[157] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer, Adversarial patch, ArXiv (2017) arXiv:1712.09665.

[158] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, Dacheng Tao, Perceptual-sensitive gan for generating adversarial patches, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (01) 2019, pp. 1028–1035.

[159] Simen Thys, Wiebe Van Ranst, Toon Goedemé, Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[160] Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, Hairong Qi, Advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8341–8350.

[161] Mikhail Pautov, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev, Aleksandr Petiushko, On adversarial patches: real-world attack on arcface-100 face recognition system, in: 2019 International Multi-Conference on Engineering, Computer and Information Sciences, (SIBIRCON), IEEE, 2019, pp. 0391–0396.

[162] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, Hang Yu, Bias-based universal adversarial patch attack for automatic check-out, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer, 2020, pp. 395–410.

[163] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, Xiangyu Zhang, Physical attack on monocular depth estimation with optimal adversarial patches, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII, Springer, 2022, pp. 514–532.

[164] Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, Tom Goldstein, Making an invisibility cloak: Real world adversarial attacks on object detectors, in: European Conference on Computer Vision, 2019.

[165] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, Xue Lin, Adversarial T-shirt! evading person detectors in a physical world, in: European Conference on Computer Vision, 2019.

[166] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.

[167] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, Cong Liu, Adv-makeup: A new imperceptible and transferable attack on face recognition, in: International Joint Conference on Artificial Intelligence, 2021.

[168] Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok, Synthesizing robust adversarial examples, in: International Conference on Machine Learning, PMLR, 2018, pp. 284–293.

[169] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, Yun Yang, Adversarial laser beam: Effective physical-world attack to DNNs in a blink, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16062–16071.

[170] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, Ivan Martinovic, SLAP: Improving physical adversarial examples with short-lived adversarial perturbations, in: 30th USENIX Security Symposium, USENIX Security 21), USENIX Association, 2021, pp. 1865–1882.

[171] Nils Worzyk, Hendrik Kahlen, Oliver Kramer, Physical adversarial attacks by projecting perturbations, in: Igor V. Tetko, Věra Kůrková, Pavel Karpov, Fabian Theis (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing, Springer International Publishing, Cham, 2019, pp. 649–659.

[172] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, Xiaolin Hu, Fooling thermal infrared pedestrian detectors in real world using small bulbs, in: AAAI Conference on Artificial Intelligence, 2021.

[173] Yanjie Li, Yiquan Li, Xuelong Dai, Songtao Guo, Bin Xiao, Physical-world optical adversarial attacks on 3D face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), 2023, pp. 24699–24708.

[174] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, Xiangyang Ji, Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), 2022, pp. 15345–15354.

[175] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, Earlence Fernandes, Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), 2021, pp. 14666–14675.

[176] Chengyu Wang, Jia Wang, Qiuzhen Lin, Adversarial attacks and defenses in deep learning: A survey, in: Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part I 17, Springer, 2021, pp. 450–461.

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

[178] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, Bastian Bischoff, On detecting adversarial perturbations, 2017, arXiv preprint arXiv:1702.04267.

[179] Jiajun Lu, Theerasit Issaranon, David Forsyth, Safetynet: Detecting and rejecting adversarial examples robustly, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 446–454.

[180] Xin Li, Fuxin Li, Adversarial examples detection in deep networks with convolutional filter statistics, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5764–5772.

[181] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, Patrick McDaniel, On the (statistical) detection of adversarial examples, 2017, arXiv preprint arXiv:1702.06280.

[182] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang, Radha Poovendran, Blocking transferability of adversarial examples in black-box learning systems, 2017, arXiv preprint arXiv:1703.04318.

[183] Dongyu Meng, Hao Chen, Magnet: A two-pronged defense against adversarial examples, in: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 135–147.

[184] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, Xiaofeng Wang, Detecting adversarial image examples in deep neural networks with adaptive noise reduction, IEEE Trans. Dependable Secure Comput. 18 (1) (2021) 72–85.

[185] Thomas Gebhart, Paul Schrater, Adversary detection in neural networks via persistent homology, 2017, arXiv preprint arXiv:1711.10056.

[186] Weilin Xu, David Evans, Yanjun Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, Proceedings 2018 Network and Distributed System Security Symposium (2018).

[187] Naveed Akhtar, Jian Liu, Ajmal Mian, Defense against universal adversarial perturbations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3389–3398.

[188] Seok-Hwan Choi, Jinmyeong Shin, Yoon-Ho Choi, PIHA: Detection method using perceptual image hashing against query-based adversarial attacks, Future Gener. Comput. Syst. (2023).

[189] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, Qi Zhao, Foveation-based mechanisms alleviate adversarial examples, 2015, arXiv preprint arXiv:1511.06292.

[190] Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G Ororbia II, Xinyu Xing, Xue Liu, C Lee Giles, Learning adversary-resistant deep neural networks, 2016, arXiv preprint arXiv:1612.01401.

[191] Gintare Karolina Dziugaite, Zoubin Ghahramani, Daniel M. Roy, A study of the effect of JPG compression on adversarial images, 2016, arXiv preprint arXiv:1608.00853.

[192] Chuan Guo, Mayank Rana, Moustapha Cisse, Laurens Van Der Maaten, Countering adversarial images using input transformations, 2017, arXiv preprint arXiv:1711.00117.

[193] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, Duen Horng Chau, Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression, ArXiv (2017) arXiv:1705.02900.

[194] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, Prateek Mittal, Enhancing robustness of machine learning systems via data transformations, in: 2018 52nd Annual Conference on Information Sciences and Systems, (CISS), IEEE, 2018, pp. 1–5.

[195] Shiwei Shen, Guoqing Jin, Ke Gao, Yongdong Zhang, APE-GAN: Adversarial perturbation elimination with gan, 2017, arXiv preprint arXiv:1707.05474.

[196] Valentina Zantedeschi, Maria-Irina Nicolae, Ambrish Rawat, Efficient defenses against adversarial attacks, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 39–49.

[197] Jonghoon Jin, Aysegul Dundar, Eugenio Culurciello, Robust convolutional neural networks under adversarial noise, 2015, arXiv preprint arXiv:1511.06306.

[198] Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G Ororbia, Xinyu Xing, Xue Liu, C Lee Giles, Adversary resistant deep neural networks with an application to malware detection, in: Proceedings of the 23rd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1145–1153.

[199] Qinglong Wang, Wenbo Guo, Alexander G Ororbia II, Xinyu Xing, Lin Lin, C Lee Giles, Xue Liu, Peng Liu, Gang Xiong, Using non-invertible data transformations to build adversarial-robust neural networks, 2016, arXiv preprint arXiv:1610.01934.

[200] Zhun Sun, Mete Ozay, Takayuki Okatani, HyperNetworks with statistical filtering for defending adversarial examples, 2017, arXiv preprint arXiv:1711.01791.

[201] Swami Sankaranarayanan, Arpit Jain, Rama Chellappa, Ser Nam Lim, Regularizing deep networks using efficient layerwise adversarial training, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, (1) 2018.

[202] Taesik Na, Jong Hwan Ko, Saibal Mukhopadhyay, Cascade adversarial machine learning regularized with a unified embedding, 2017, arXiv preprint arXiv:1708.02582.

[203] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, Tom Goldstein, Adversarial training for free!, in: Advances in Neural Information Processing Systems, 2019, pp. 3358–3369.

[204] Takeru Miyato, Andrew M. Dai, Ian Goodfellow, Adversarial training methods for semi-supervised text classification, 2016, arXiv preprint arXiv:1605.07725.

[205] Stephan Zheng, Yang Song, Thomas Leung, Ian Goodfellow, Improving the robustness of deep neural networks via stability training, in: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, 2016, pp. 4480–4488.

[206] Hyeungill Lee, Sungyeob Han, Jungwoo Lee, Generative adversarial trainer: Defense to adversarial perturbations with gan, 2017, arXiv preprint arXiv:1705.03387.

[207] Qizhang Li, Yiwen Guo, Wangmeng Zuo, Hao Chen, Squeeze training for adversarial robustness, in: The Eleventh International Conference on Learning Representations, 2023.

[208] Chunchuan Lyu, Kaizhu Huang, Hai-Ning Liang, A unified gradient regularization family for adversarial examples, in: 2015 IEEE International Conference on Data Mining, IEEE, 2015, pp. 301–309.

[209] Uri Shaham, Yutaro Yamada, Sahand Negahban, Understanding adversarial training: Increasing local stability of neural nets through robust optimization, 2015, arXiv preprint arXiv:1511.05432.

[210] Andrew Ross, Finale Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, (1) 2018.

[211] Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi, Deepcloak: Masking deep neural network models for robustness against adversarial samples, 2017, arXiv preprint arXiv:1702.06763.

[212] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: 2016 IEEE Symposium on Security and Privacy, (SP), IEEE, 2016, pp. 582–597.

[213] Nicolas Papernot, Patrick McDaniel, On the effectiveness of defensive distillation, 2016, arXiv preprint arXiv:1607.05113.

[214] Nicolas Papernot, Patrick McDaniel, Extending defensive distillation, 2017, arXiv preprint arXiv:1705.05264.

[215] Varun Chandrasekaran, Brian Tang, Nicolas Papernot, Kassem Fawaz, Somesh Jha, Xi Wu, Rearchitecting classification frameworks for increased robustness, 2019, arXiv preprint arXiv:1905.10900.

[216] Scott Freitas, Shang-Tse Chen, Zijie J. Wang, Duen Horng Chau, Unmask: Adversarial detection and defense through robust feature alignment, in: 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 1081–1088.

[217] Xiao Li, Ziqi Wang, Bo Zhang, Fuchun Sun, Xiaolin Hu, Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 45 (7) (2023) 8861–8873.

[218] Chawin Sitawarin, Kornrapat Pongmala, Yizheng Chen, Nicholas Carlini, David Wagner, Part-based models improve adversarial robustness, in: The Eleventh International Conference on Learning Representations, 2023.

[219] Jin Ding, Jie-Chao Zhao, Yong-Zhi Sun, Ping Tan, Ji-En Ma, You-Tong Fang, Improving the robustness of deep convolutional neural networks through feature learning, 2023, arXiv preprint arXiv:2303.06425.

[220] Taehoon Lee, Minsuk Choi, Sungroh Yoon, Manifold regularized deep neural networks using adversarial examples, 2015, arXiv preprint arXiv:1511.06381.

[221] Thilo Strauss, Markus Hanselmann, Andrej Junginger, Holger Ulmer, Ensemble methods as a defense to adversarial perturbations against deep neural networks, 2017, arXiv preprint arXiv:1709.03423.

[222] Navid Kardan, Kenneth O. Stanley, Mitigating fooling with competitive overcomplete output layer neural networks, in: 2017 International Joint Conference on Neural Networks, (IJCNN), IEEE, 2017, pp. 518–525.

[223] Moustapha Cisse, Yossi Adi, Natalia Neverova, Joseph Keshet, Houdini: Fooling deep structured prediction models, 2017, arXiv preprint arXiv:1707.05373.

[224] Linh Nguyen, Sky Wang, Arunesh Sinha, A learning and masking approach to secure learning, in: Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings 9, Springer, 2018, pp. 453–464.

[225] Jia Liu, Yaochu Jin, Evolving hyperparameters for training deep neural networks against adversarial attacks, in: 2019 IEEE Symposium Series on Computational Intelligence, (SSCI), 2019, pp. 1778–1785.

[226] Yigit Alparslan, Edward Kim, ATRAS: Adversarially trained robust architecture search, 2021, arXiv preprint arXiv:2106.06917.

[227] Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, Xingjun Ma, Exploring architectural ingredients of adversarially robust deep neural networks, Adv. Neural Inf. Process. Syst. 34 (2021) 5545–5559.

[228] Nanqing Dong, Min Xu, Xiaodan Liang, Yiliang Jiang, Wei Dai, Eric Xing, Neural architecture search for adversarial medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 828–836.

[229] Ramtin Hosseini, Xingyi Yang, Pengtao Xie, DSRNA: Differentiable search of robust neural architectures, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 6196–6205.

[230] Jisoo Mok, Byunggook Na, Hyeokjun Choe, Sungroh Yoon, AdvRush: Searching for adversarially robust neural architectures, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12322–12332.

[231] Ke Wang, Peng Xu, Chien-Ming Chen, Saru Kumari, Mohammad Shojafar, Mamoun Alazab, Neural architecture search for robust networks in 6G-enabled massive IoT domain, IEEE Internet Things J. 8 (7) (2021) 5332–5339.

[232] Danilo Vasconcellos Vargas, Shashank Kotyan, Evolving robust neural architectures to defend from adversarial attacks, 2019, arXiv preprint arXiv:1906.11667.

[233] Guoyang Xie, Jinbao Wang, Guo Yu, Feng Zheng, Yaochu Jin, Tiny adversarial mulit-objective oneshot neural architecture search, 2021, arXiv preprint arXiv:2103.00363.

[234] Jia Liu, Yaochu Jin, Multi-objective search of robust neural architectures against multiple types of adversarial attacks, Neurocomputing 453 (2021) 73–84.

[235] Jia Liu, Ran Cheng, Yaochu Jin, Bi-fidelity evolutionary multiobjective search for adversarially robust deep neural architectures, 2022, arXiv preprint arXiv:2207.05321.

[236] Zhixiong Yue, Baijiong Lin, Xiaonan Huang, Yu Zhang, Effective, efficient and robust neural architecture search, 2020, arXiv preprint arXiv:2011.09820.

[237] Xuefei Ning, Junbo Zhao, Wenshuo Li, Tianchen Zhao, Yin Zheng, Huazhong Yang, Yu Wang, Discovering robust convolutional architecture at targeted capacity: A multi-shot approach, 2020, arXiv preprint arXiv:2012.11835.

[238] Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, Dahua Lin, When NAS meets robustness: In search of robust architectures against adversarial attacks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 631–640.

[239] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, Vineeth N Balasubramanian, On adversarial robustness: A neural architecture search perspective, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 152–161.

[240] Hanlin Chen, Baochang Zhang, Song Xue, Xuan Gong, Hong Liu, Rongrong Ji, David Doermann, Anti-bandit neural architecture search for model defense, in: European Conference on Computer Vision, Springer, 2020, pp. 70–85.

[241] George Cazenavette, Calvin Murdock, Simon Lucey, Architectural adversarial robustness: The case for deep pursuit, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), 2021, pp. 7150–7158.

[242] Nicholas Carlini, David Wagner, Magnet and "efficient defenses against adversarial attacks" are not robust to adversarial examples, 2017, arXiv preprint arXiv:1711.08478.

[243] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, Andrew B Gardner, Detecting adversarial samples from artifacts, 2017, arXiv preprint arXiv:1703.00410.

[244] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[245] Richard Shin, Dawn Song, Jpeg-resistant adversarial images, in: NIPS 2017 Workshop on Machine Learning and Computer Security, Vol. 1, 2017, p. 8.

[246] Weilin Xu, David Evans, Yanjun Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, 2017, arXiv preprint arXiv:1704.01155.

[247] Shan Sung Liew, Mohamed Khalil-Hani, Rabia Bakhteri, Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems, Neurocomputing 216 (2016) 718–734.

[248] Harris Drucker, Yann Le Cun, Improving generalization performance using double backpropagation, IEEE Trans. Neural Netw. 3 (6) (1992) 991–997.

[249] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[250] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, Yoshua Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, 2011, pp. 833–840.

[251] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, TAMT Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.

[252] Hanxiao Liu, Karen Simonyan, Yiming Yang, DARTS: Differentiable architecture search, in: International Conference on Learning Representations, 2019.

[253] Esteban Real, Alok Aggarwal, Yanping Huang, Quoc V. Le, Regularized evolution for image classifier architecture search, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (01) 2019, pp. 4780–4789.

[254] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song, Natural adversarial examples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15262–15271.

[255] Dan Hendrycks, Thomas Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: Proceedings of the International Conference on Learning Representations, 2019.

[256] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al., The many faces of robustness: A critical analysis of out-of-distribution generalization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8340–8349.

[257] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, Hui Xue, ImageNet-E: Benchmarking neural network robustness via attribute editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20371–20381.

[258] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, Zhang Huan, Advbox: a toolbox to generate adversarial examples that fool neural networks, 2020.

[259] Gavin Weiguang Ding, Luyu Wang, Xiaomeng Jin, AdverTorch v0.1: An adversarial robustness toolbox based on pytorch, 2019, arXiv preprint arXiv:1902.07623.

[260] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, Shibao Zheng, A comprehensive study on robustness of image classification models: Benchmarking and rethinking, 2023, arXiv preprint arXiv:2302.14301.

[261] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, Ben Edwards, Adversarial robustness toolbox v1.2.0, CoRR 1807.01069 (2018).

[262] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, Technical report on the CleverHans v2.1.0 adversarial examples library, 2018, arXiv preprint arXiv:1610.00768.

[263] Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, Ting Wang, Deepsec: A uniform platform for security analysis of deep learning model, in: 2019 IEEE Symposium on Security and Privacy, (SP), IEEE, 2019, pp. 673–690.

[264] Jonas Rauber, Roland Zimmermann, Matthias Bethge, Wieland Brendel, Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, TensorFlow, and JAX, J. Open Source Softw. 5 (53) (2020) 2607.

[265] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H.S. Torr, Dacheng Tao, Robustart: Benchmarking robustness on architecture design and training techniques, 2021, https://arxiv.org/pdf/2109.05211.pdf.

[266] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, Matthias Hein, RobustBench: A standardized adversarial robustness benchmark, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

[267] Steffen Jung, Jovita Lukasik, Margret Keuper, Neural architecture design and robustness: A dataset, in: The Eleventh International Conference on Learning Representations, 2023.

[268] Preetum Nakkiran, Adversarial robustness may be at odds with simplicity, 2019, arXiv preprint arXiv:1901.00532.

[269] Amirata Ghorbani, Abubakar Abid, James Zou, Interpretation of neural networks is fragile, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (01) 2019, pp. 3681–3688.

[270] Liwei Song, Reza Shokri, Prateek Mittal, Privacy risks of securing machine learning models against adversarial examples, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 241–257.

Jia Liu received the B.Sc. degree from Shenyang University of Technology, Shenyang, China, in 2015 and the M.Sc. degree from Shenzhen University, Shenzhen, China, in 2018. She received the Ph.D. degree from University of Surrey, Guildford, United Kingdom, in 2022. She is currently a Computer Vision Algorithm Engineer with Ping An Property & Casualty Insurance Company, Shenzhen, China. Her research interests include computer vision, adversarial deep learning, and evolutionary neural architecture search.

Yaochu Jin obtained the BSc., MSc. and Ph.D. degree all in automatic control from the Electrical Engineering Department, Zhejiang University, China, in 1988, 1991, and 1996, respectively, and the Dr.-Ing. from the Institute of Neuroinformatics, Ruhr-University Bochum, Germany in 2001. He is currently a Chair Professor of Artificial Intelligence with the School of Engineering, Westlake University. Before joining Westlake University, he was an Alexander von Humboldt Professor for Artificial Intelligence endowed by the German Federal Ministry of Education and Research, Bielefeld University, Germany from 2021 to 2023, and a Surrey Distinguished Chair Professor in Computational Intelligence, University of Surrey, Guildford, U.K., from 2010 to 2021. He was a "Finland Distinguished Professor" of University of Jyväskylä, Finland, and "Changjiang Distinguished Visiting Professor", Northeastern University, China from 2015 to 2017. Prof Jin is presently the President-Elect of the IEEE Computational Intelligence Society and the Editor-in-Chief of Complex & Intelligent Systems. He was the Editor-in-Chief of the IEEE Transactions on Cognitive and Developmental Systems, an IEEE Distinguished Lecturer in 2013-2015 and 2017-2019, the Vice President for Technical Activities of the IEEE Computational Intelligence Society (2015-2016). He is the recipient of the 2018, 2021 and 2023 IEEE Transactions on Evolutionary Computation Outstanding Paper Award, and the 2015, 2017, and 2020 IEEE Computational Intelligence Magazine Outstanding Paper Award. He was named as a "Highly Cited Researcher" consecutively from 2019 to 2023 by Clarivate. He is a Member of Academia Europaea and Fellow of IEEE.