Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey

Yulong Wang, *Member, IEEE*, Tong Sun, Shenghong Li, *Member, IEEE*, Xin Yuan, *Member, IEEE*, Wei Ni, *Senior Member, IEEE*, Ekram Hossain, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—Adversarial attacks and defenses in machine learning and deep neural network (DNN) have been gaining significant attention due to the rapidly growing applications of deep learning in communication networks. This survey provides a comprehensive overview of the recent advancements in the field of adversarial attack and defense techniques, with a focus on DNN-based classification models for communication applications. Specifically, we conduct a comprehensive classification of recent adversarial attack methods and state-of-the-art adversarial defense techniques based on attack principles, and present them in visually appealing tables and tree diagrams. This is based on a rigorous evaluation of the existing works, including an analysis of their strengths and limitations. We also categorize the methods into counter-attack detection and robustness enhancement, with a specific focus on regularization-based methods for enhancing robustness. New avenues of attack are also explored, including search-based, decision-based, drop-based, and physical-world attacks, and a hierarchical classification of the latest defense methods is provided, highlighting the challenges of balancing training costs with performance, maintaining clean accuracy, overcoming the effect of gradient masking, and ensuring method transferability. At last, the lessons learned and open challenges are summarized with future research opportunities recommended.**

***Index Terms*—Machine learning, deep neural network, adversarial attack, adversarial defense, communication, network.**

Manuscript received 10 March 2023; revised 30 July 2023; accepted 20 September 2023. Date of publication 26 September 2023; date of current version 22 November 2023. This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61921003; in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery Grant RGPIN-2019-05870; and in part by the U.S. National Science Foundation under Grant CNS-2128448 and Grant ECCS-2335876. *(Corresponding author: Yulong Wang.)*

Yulong Wang and Tong Sun are with the State Key Laboratory of Networking and Switching Technology, School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: wyl@bupt.edu.cn; suntong@bupt.edu.cn).

Shenghong Li, Xin Yuan, and Wei Ni are with the Data61, Commonwealth Science and Industrial Research Organisation, Sydney, NSW 2122, Australia (e-mail: shenghong.li@data61.csiro.au; xin.yuan@data61.csiro.au; wei.ni@data61.csiro.au).

Ekram Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada (e-mail: ekram.hossain@umanitoba.ca).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/COMST.2023.3319492

Abbreviation	*Full form*
AAAM	Adversarial Attack Attention Module
adMRL	adversarial MRL
AdvCam	Adversarial Camouflage
AdvLB	Adversarial Laser Beam
AdvRush	Adversarially robust architecture Rush
AGKD-BML	Attention Guided Knowledge Distillation and Bi-directional Metric Learning
AI	Artificial Intelligence
AL	Adversarial Learning
AMGmal	Adaptive Mask-Guided adversarial attack against malware detection
AMM	Adversarial Margin Maximization Network
AMC	Automatic Modulation Classification
ANP	Adversarial Noise Propagation
AoA	Attack on Attention
APE-GAN	Adversarial Perturbation Elimination GAN
APR	Amplitude-Phase Recombination
ART	Adaptive ReTraining
ASR	Attack Success Rate
AT	Adversarial Training
AtkSE	Attacking by Shrinking Error
ATTA	Adversarial Transformation-enhanced Transfer Attack
AUC	Area Under Curve
BER	Bit Error Rate
BN	Batch Normalization
BIM	Basic Iterative Method
BLF	Bounded Logit Function
BFGS	Broyden Fletcher Goldfarb Shanno
CAFA	Class Activation Feature Loss
CAP-GAN	Cycle-consistent Attentional Purification GAN
C-BCE	Conditional Binary Cross-Entropy
CD	Constellation Diagram
CE	Cross-Entropy
CAFD	Class Activation Feature-based Denoiser
C-LSTM	Conv-Long Short-Term Memory
CMAG	Cascade Model-Aware Generative
CNN	Convolutional Neural Network
COLT	COnvex Layer-wise Adversarial Training
CSA	Cost-Sensitive Adversarial learning model
CSE	Cost-Sensitive Adversarial Extension
CSM	Cross-Spectral Mapping

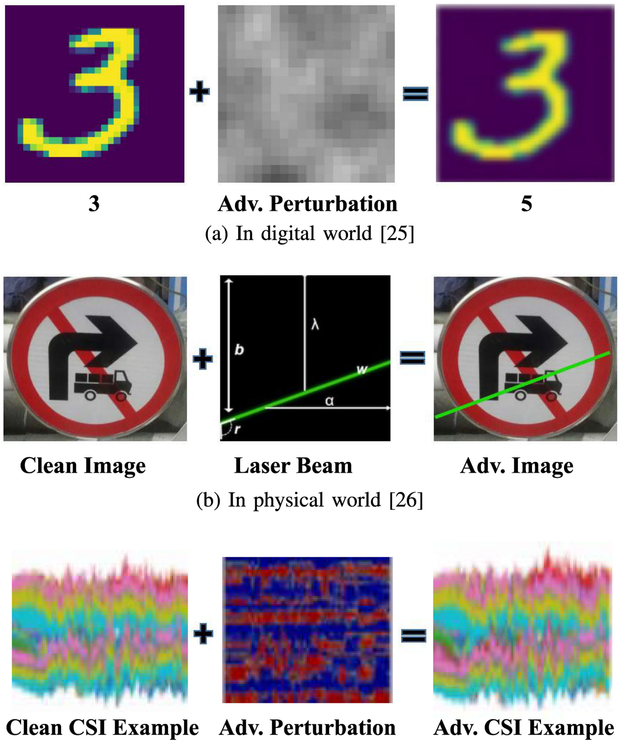
CSI	Channel State Information
C&W	Carlini and Wagne attacking algorithm
DAC	Degree Assortativity Change
DGA	Domain Generating Algorithm
DH-AT	Dual Head Adversarial Training
DL	Deep Learning
DM	Data Manifold
DNN	Deep Neural Network
DSNGD	Dynamically Sampled Nonlocal Gradient Descent
D2Defend	Dual-Domain based Defense
DWT	Discrete Wavelet Transform
ER-Classifier	Embedding Regularized Classifier
ERF	Effective Receptive Field
FeaCP	Feature-wise Convex Polytope attack
FGMD	Feature Grouping and Multi-model Fusion Detector
FGSM	Fast Gradient Sign Method
FMR	Feature-Map Reconstructor
FNC	Feature Norm Clipping
FP	False Positive
FR	Face Recognition
FS	Feature Scattering
GAE	Graph AutoEncoder
GAN	Generative Adversarial Network
GAP	Generation Adversarial Perturbation
GAT	Generative Adversarial Training
GCN	Graph Convolutional Network
GEM	Graph Embedding Model
GF-Attack	Generalized adversarial attack Framework
GNN	Graphic Neural Network
GPMT	Generating Practical Malicious Traffic
GR	Global Reconstructor
GraphAT	Graph Adversarial Training
H&G	Hendrycks and Gimpel
HAG	Hash Adversary Generation
HFT	High Frequency Trading
HMC	Hamiltonian Monte Carlo
HMCAM	Hamiltonian Monte Carlo with Accumulated Momentum
HPGD	Hybrid Projection Gradient Descent
HSI	Hyper-Spectral Image
IBA	Iterative Black-box Attack
ICAT	Induced Class Adversarial Training
IDS	Intrusion Detection System
IGA	Iterative Gradient Attack
IIoT	Industrial Internet of Things
IoT	Internet of Things
IoU	Intersection over Union
IoV	Internet of Vehicle
IPW	Iterative Partially-White-box subspace attack
JCR	Journal Citation Reports
JSMA	Jacobian-based Saliency Map Attack
KD	Knowledge Distillation
KL	Kullback Leibler divergence
KR	Kantorovich-Rubinstein
LAFEAT	LAtent FEAture Attack
L-BFGS	Limited-memory BFGS
LF	Low-Frequency component distortion
LID	Local Intrinsic Dimensionality
LPIPS	Learned Perceptual Image Patch Similarity
LR	Logit Reconstructor
LS-GNA	Label Smoothing and Gaussian Noise Augmentation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAP	Multispectral Adversarial Patch
ME	Material Emissivity
MI-FGSM	Momentum Iterative Fast Gradient Sign Method
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MLP	MuLtilayer Perceptron
MPA	Most Powerful Attack
MRL	Meta Reinforcement Learning
MultiD-WGAN	Muti-Discriminator Wasserstein GAN
N-BaIoT	Network-Based detection of IoT
NAS	Neural Architecture Search
NAttack	NES adversarial Attack
NES	Natural Evolution Strategy
NID	Network Intrusion Detection
NIDS	Network Intrusion Detection System
NLM	Non-Local spatial smoothing
NLP	Natural Language Processing
NR	Neural Rejection
NSGA-PSO	Non-dominated Sorting Genetic Algorithm with Particle Swarm Optimization
NSS	Normalized Scanpath Saliency (minus)
ODE	Ordinary Differential Equation
OFDM	Orthogonal Frequency Division Multiplexing
OOD	Out Of Domain inputs
PCA	Principal Component Analysis
PCAE	Principal Component Adversarial Example
PCL	Prototype Conformity Loss
PDA	Progressive Diversified Augmentation
PEC	Polyhedral Envelope Certifier
PER	Polyhedral Envelope Regularization
PGD	Projected Gradient Descent
PLC	Pearson's Linear Coefficient (minus)
PR	Precision-Recall
PS-GAN	Perceptual-Sensitive GAN
PSO	Particle Swarm Optimization
PUAA	Primary User Adversarial Attack
QoS	Quality of Service
RE	Reconstruction Error
RF	Radio Frequency
RFFI	Radio Frequency FIngerprinting
RISs	Reconfigurable Intelligent Surfaces
RMS	Root Mean Square
ROC	Receiver Operating characteristic Curve
ROSA	RObust SAliency
RP2	Robust Physical Perturbation
RRF	Rectified Reverse Function
RSLAD	Robust Soft Label Adversarial Distillation
RMSE	Root Mean Squared Error
SACNet	Self-Attention Context Network

SAD	Saliency Adversarial Defense
SCA	Side Channel Attack
SCE	Softmax Cross-Entropy
SD-Alg	Shortest Distance Algorithm
SGA	Simplified Gradient-based Attack
SGADV	Similarity-based Gray-box Adversarial Attack
SML	Single-directional Metric Learning
SNN	Spiking Neural Network
SNS	Sensitive Neuron Stabilizing
SOTA	State Of The Art
SPSA	Simultaneous Perturbation Stochastic Approximation
SRLIM	Surrogate Representation Learning with Isometric Mapping
SSAH	Semantic Similarity Attack on High-frequency components
SSIM	Structural Similarity
STDB	Spike-Timing-Dependent Backpropagation
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TAD	Transfer learning-based multi-Adversarial Detection
TP	True Positive
TriATNE	Tripartite Adversarial Training for Network Embeddings
TTA	Traffic Type Analysis
UAP	Universal Adversarial Perturbation
VAM	Virtual Adversarial Method
VGG	Visual Geometry Group
WaveCNet	Wavelet-integrated Convolutional Network
WD	Wasserstein Distance
ZOO	Zero-Order Optimization

I. Introduction

DEEP neural networks (DNNs) are a crucial component of the artificial intelligence (AI) landscape due to their ability to perform complex tasks, modulation recognition [1], [2], wireless signal classification [3], [4], network intrusion detection and defense [5], [6], [7], object detection [8], [9], object tracking [10], [11], image classification [12], [13], [14], language translation [15], [16], and many more [17], [18], [19]. The availability of advanced hardware, such as GPUs, TPUs, and NPUs, has facilitated the training of DNNs and made them a popular research direction in AI [20], [21]. However, despite their strong learning ability, DNNs are susceptible to adversarial attacks, such as classical attack method Projected Gradient Descent (PGD) [22], Square attack [23] or C&W [24]. These attacks exploit the model's sensitivity to small and carefully crafted perturbations in the input data, causing the DNN to produce false predictions. Adversarial attacks represent a serious challenge to the robustness of DNNs and require proactive attention and action to mitigate the risks they pose.

Adversarial attacks in DNN can have serious consequences, as captured in many recent studies. For example, Fig. 1 also illustrates various such attacks, including a deliberately devised alteration to an input image resulting in misclassification by a convolutional neural network (CNN) with a 99% level of certainty [25], a traffic sign recognition attack that uses a laser beam to fool self-driving cars [26], and a channel state information (CSI) recognition attack in an Internet of Things (IoT) scenario, where CSI examples are adversarially perturbed to mislead DNN models [27]. Within the area of signal recognition, the construction of adversarial samples can be achieved by incorporating disturbance noise into signal waveforms [28]. This manipulation can lead to erroneous predictions or recognition by machine learning (ML) models. Such discrepancies can pose significant threats to contemporary wireless communication systems, cognitive wireless networks, satellite navigation, and electromagnetic reconnaissance. To this end, adversarial attacks can pose significant risks and impacts on many important areas with ML or DNN involved, especially on communications and networking, as articulated in the following.



(a) In digital world [25]

(b) In physical world [26]

(c) In cyberspace [27]

Fig. 1. Examples of adversarial perturbation.

A. Impacted Areas

Adversarial attacks, which add imperceptible disturbances to input samples, emerged in computer vision and can deceive systems such as face recognition [29], [30], [31], object detection [32], [33], and object tracking [34], [35]. They now pose risks across various fields, including autonomous driving [26], [36], [37], [38], [39], finance [40], [41], and human-machine interaction [42], [43]. With the widespread adoption of DNN models in the communication and network domain, the impact of adversarial attacks cannot be overlooked. The areas within communication and networking impacted by adversarial attacks include signal processing, network security, network management, and resource allocation.

1) Signal Processing: To facilitate signal detection and fast tracking, DNN-based wireless signal classifiers have been utilized by wireless signal receivers to categorize over-the-air received signals into different modulation schemes and orders. It is demonstrated in [44] that these DNN models are vulnerable to channel-aware adversarial attacks. In such attacks, an adversary can transmit an adversarial perturbation within a given power budget to mislead the receiver into making incorrect predictions when classifying wireless signals superimposed with the adversarial perturbation. Adversarial attacks can also substantially degrade the performance of DNN models used for modulation scheme recognition in communication systems [1], [2], [45]. The authors of [46] indicate that adversarial attacks significantly increase the bit error rate (BER) of a DNN-based multi-user orthogonal frequency-division multiplexing (OFDM) detectors, which are trained to recover the payload bits directly from received symbols. In a spectrum monitoring scenario, adversarial attacks in the form of adversarial waveforms can successfully disrupt attempts to intercept and classify signals using convolutional neural networks (CNNs), and the attacking success rate increases as bandwidth increases [28], [47]. In addition, Generative Adversarial Networks (GANs), previously used for generating synthetic image examples [48], have now been adapted to produce adversarial examples that attack modulation classifiers in wireless communications [3], [4], [49] or conduct wireless signal spoofing [50], [51]. Signals can be used to train ML models for identifying IoT devices, which may also be threatened by adversarial attacks capable of constructing specially spoofed signals [52].

Moreover, Huang et al. [53] show that adversarially contaminated Wi-Fi signals could mislead DNN-based non-intrusive human activity recognition systems. Xu et al. [27] construct adversarial perturbation by customizing Fast Gradient Sign Method (FGSM) [54] and PGD [22], and reduce the performance of Wi-Fi sensing applications, such as user identification [55], gesture recognition, and human activity recognition [56]. With tiny perturbation-to-signal ratios of around −18 dB in CSI-based Wi-Fi fingerprinting, adversarial attacks can reach an extraordinary attack success rate of over 90% [57].

At the same time, numerous defensive strategies have been developed to counteract adversarial instances in radio signals modulation [58], signal classification [59], [60], and modulation classification [2], [61], [62], [63], [64]. Predominantly, the primary research methodology strives to minimize the attack surface and sustain automatic modulation classification (AMC) [2], even in the face of meticulously crafted adversarial attacks.

2) Resource Allocation: In addition to classification tasks, adversarial attacks affect regression problems that can significantly damage the power allocation process in massive multiple-input multiple-output (MIMO) networks, often leading to unfeasible solutions [65], [66], [67]. For instance, Boora et al. [66] use CNN and ordinary differential equation (ODE) models to study the effects of adversarial attack and defense on massive MIMO localization, and verify that adversarial training-based neural ODE can effectively improve the robustness of massive MIMO localization in indoor and outdoor environments. Manoj et al. [67] study adversarial attacks against DNN-based optimal power allocation in a massive MIMO system, demonstrating that adding only a small perturbation to the input of DNNs can lead to a strong attack consequence.

3) Network Management: Adversarial attacks have also been observed in network management, due to the increasing application of DNN-based network traffic classification [68], [69]. Universal Adversarial Perturbation (UAP) [70], originally developed for adversarial attacks against DNN-based image classifiers, has been evaluated for attacking DNN-based network traffic classification [71]. Nowroozi et al. [72] show that adversarial attacks, such as Jacobian-based Saliency Map (JSMA) [73], Iterative Fast Gradient Method (I-FGSM) [74], PGD [22], Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS) [75], and DeepFool attack [76], can be used to attack CNN models trained on well-known computer network datasets, including the Domain Generating Algorithms (DGA) dataset, the Network-based Detection of IoT (N-BaIoT) dataset, and the RIPE Atlas dataset, with attack success rates ranging from 63% to 100%. DNN models trained using collected TCP/IP traffic data can be fooled by perturbed network packets sent from the host controlled by an attacker [77], [78].

There is a burgeoning interest in adversarial defense schemes for network traffic classification, due to the significant threats instigated by DNN-based packet sniffers. For an instance, Yang et al. [79] propose an effective traffic obfuscation method based on neural networks, which generates traffic distortions with minimal overhead and computational cost but attains comparable obfuscation performance. Such obfuscation can effectively defend eavesdropping or traffic analysis attacks.

4) Network Intrusion Detection: Network Intrusion Detection Systems (NIDS) are extensively utilized for the detection and filtration of malicious network traffic [20], [80], [81]. ML-based detectors [82], in particular, offer an effective means of recognizing intrusive network packets. However, attackers have found ways to bypass NIDS by generating adversarial samples. This is typically accomplished by subtly altering a small subset of traffic characteristics, such as the interval between successive packets, or by introducing entirely new features until they successfully bypass the NIDS [80], [83]. Sun et al. [68] propose an adversarial attack framework to generate malicious practical traffic with little prior knowledge to deceive ML-based detection, which can be universally adapted to multiple malicious traffic. Adversarial attacks, such as transfer learning-based multi-adversarial detection (TAD) [83], customized AT [59], and an adaptive mask-guided adversarial attack against malware detection (AMGmal) [80], have been proven capable of circumventing these detection systems. Zhang et al. [5] reveal that adversarial attacks based on perturbed network traffic, can evade an NIDS with a success rate of up to 35.7%. To evaluate the risks posed by adversarial attacks in the Industrial Internet of Things (IIoT), methods

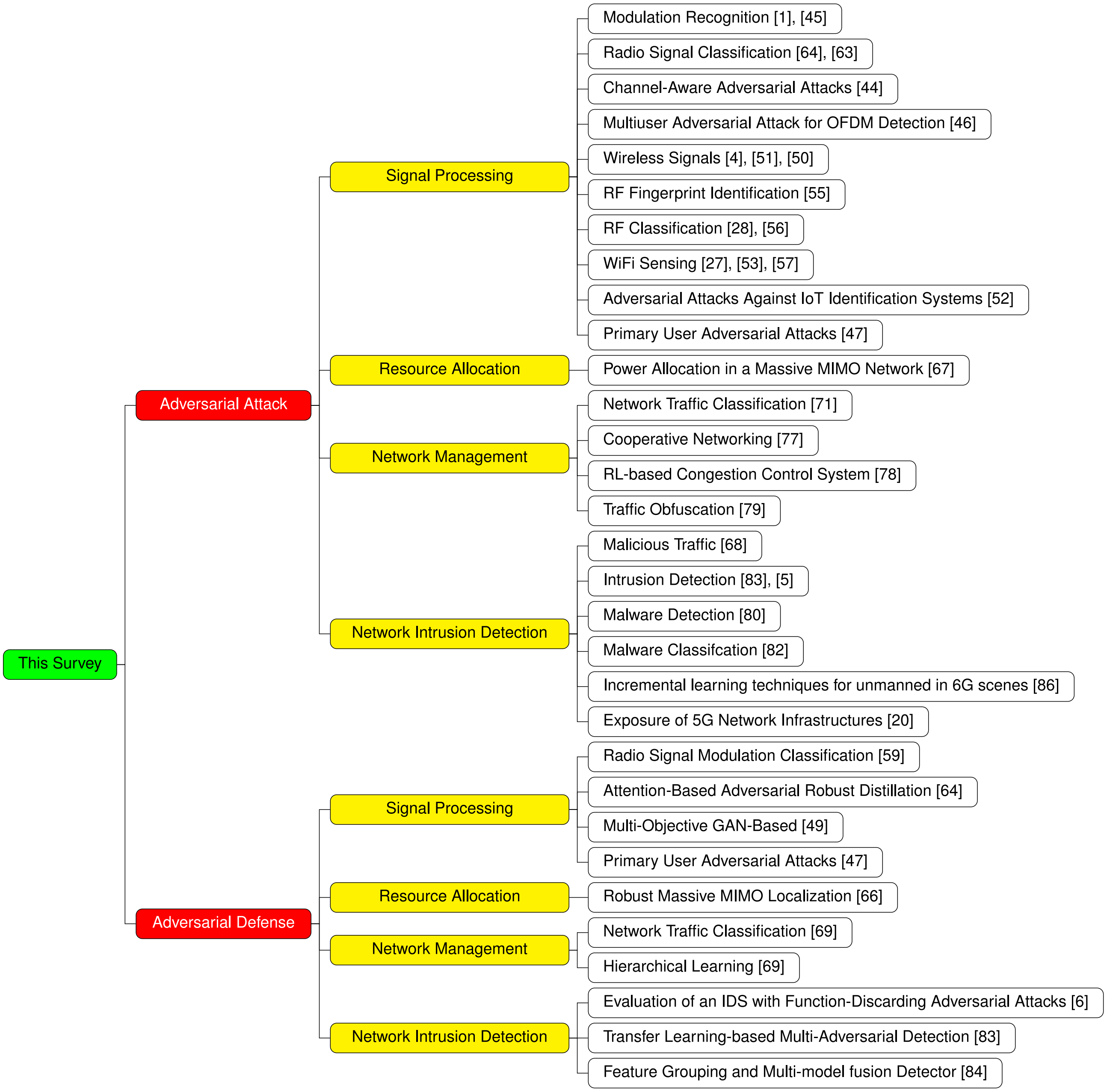

This Survey
Adversarial Attack
Signal Processing
Modulation Recognition [1], [45]
Radio Signal Classification [64], [63]
Channel-Aware Adversarial Attacks [44]
Multiuser Adversarial Attack for OFDM Detection [46]
Wireless Signals [4], [51], [50]
RF Fingerprint Identification [55]
RF Classification [28], [56]
WiFi Sensing [27], [53], [57]
Adversarial Attacks Against IoT Identification Systems [52]
Primary User Adversarial Attacks [47]
Resource Allocation
Power Allocation in a Massive MIMO Network [67]
Network Management
Network Traffic Classification [71]
Cooperative Networking [77]
RL-based Congestion Control System [78]
Traffic Obfuscation [79]
Network Intrusion Detection
Malicious Traffic [68]
Intrusion Detection [83], [5]
Malware Detection [80]
Malware Classifcation [82]
Incremental learning techniques for unmanned in 6G scenes [86]
Exposure of 5G Network Infrastructures [20]
Adversarial Defense
Signal Processing
Radio Signal Modulation Classification [59]
Attention-Based Adversarial Robust Distillation [64]
Multi-Objective GAN-Based [49]
Primary User Adversarial Attacks [47]
Resource Allocation
Robust Massive MIMO Localization [66]
Network Management
Network Traffic Classification [69]
Hierarchical Learning [69]
Network Intrusion Detection
Evaluation of an IDS with Function-Discarding Adversarial Attacks [6]
Transfer Learning-based Multi-Adversarial Detection [83]
Feature Grouping and Multi-model fusion Detector [84]

Fig. 2. Taxonomy of seminal work conducted in the field of adversarial attacks on communication networks since 2021.

like Evasion-Injection-Fabrication-Denial Adversarial Attack (EIFDAA) [6] and AMGmal [80] have been proposed. These methods aim to deceive an ML-based NIDS used within the IIoT, and to measure the performance of these ML-based NIDS systems within IIoT networks. More advanced research in this field involves assessing the defensive performance of NIDS against a variety of adversarial attack algorithms, and the design of new defense strategies [6], [83], [84].

5G networks, which are envisioned to support billions of heterogeneous devices with quality of service (QoS) provisioning, are expected to heavily rely on ML. As a result, these 5G environments will be susceptible to adversarial attacks [85]. However, due to the scarcity of ML-driven 5G devices available for adversarial ML research, proactively assessing such risks is a significant challenge. Concurrently, advancements in mobile communication, particularly the emerging ultra-low delay 6G technology, can substantially improve Internet of Vehicle (IoV) technology and enhance autonomous driving. Nevertheless, adversarial attacks present security concerns, particularly in the area of autonomous scene recognition [86]. These attacks can be exploited in-vehicle networking systems, potentially leading to traffic accidents and jeopardizing personal safety. Therefore, with the imminent advent of 6G technology, it is crucial to consider the safety implications of using deep learning (DL) algorithms in connected vehicle systems.

Moreover, we have gathered additional relevant citations and organized them based on their associations with various aspects of communications and networking. The outcomes of this effort are depicted in Fig. 2.

B. Attack Scenario

Adversarial attacks can occur during a model inference stage. Specifically, an attacker can aim to deceive a DNN-based model, e.g., a sample classifier, by launching a two-phase attack: Generating an adversarial example from the DNN-based sample classifier and feeding it back into the sample classifier. In the field of communication and networking empowered by ML models, attackers can gain knowledge about the models through several methods:

1) Model Stealing [87]: The attacker can train a surrogate model that imitates the behavior of the target model by sending queries and observing responses. This technique is also known as "model extraction".
2) Model Inversion [88]: In this scenario, the attacker attempts to reconstruct the original training data or some data properties, given a trained model and some auxiliary information.
3) Membership Inference [89]: This attack attempts to ascertain whether a particular data instance was part of the training dataset or not, thereby potentially revealing sensitive information.

A practical example might involve an attacker attempting to compromise an NIDS that employs an ML model [90]. The attacker could employ techniques like model extraction to comprehend how the NIDS categorizes normal and malicious network traffic. Once the attacker gains a solid understanding of the model's behavior, it can craft adversarial network packets that appear benign to the NIDS but are, in fact, malicious.

During the first stage, the attacker perturbs the element values of a benign example to maximize the loss function value of the sample classifier. This forces the sample classifier to misclassify the adversarially perturbed example or minimize the loss function value with regards to an incorrect class designed by the attacker. The attacker can employ different strategies to guide the direction of perturbation based on their *a-priori* knowledge of the DNN model under attack. If the neural network architecture, learned parameter values (weights and biases), and the loss function of the DNN model is available (e.g., due to a compromised server or a rogue employee), the attacker can exploit a gradient-based attacking algorithm to calculate the perturbation and produce the adversarial example.

For instance, FGSM [54] generates an adversarial instance, denoted by $\mathbf{x}^{adv}$, by applying the following rule:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \times \text{sign}(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, y)),$$

where $\mathbf{x}$ is the input data, $\epsilon \in \mathbb{R}^+$ is the perturbation magnitude, y indicates the ground-truth class, $\text{sign}(\cdot)$ returns the sign of a real value, and $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, y)$ indicates the gradient of the loss function $\mathcal{L}(\mathbf{x}, y)$ with regard to the input example $\mathbf{x}$.

PGD [22] is another typical adversarial attack algorithm, which improves FGSM by generating an adversarial example iteratively:

$$\mathbf{x}^{i+1} = \prod_{\mathbf{x}+\mathcal{S}_\epsilon} \Big(\mathbf{x}^i + \alpha \ \text{sign}(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, y))\Big), \tag{1}$$

where i is the index to an iteration, $\alpha \leq \epsilon$ is the perturbation step size, $\mathcal{S}_\epsilon \subseteq \mathbb{R}^d$ is the set of allowed perturbations under the maximum perturbation magnitude ϵ, and the projector $\prod_{\mathbf{x}+\mathcal{S}_\epsilon}(\cdot)$ maps its input to the closest element to the input in the set $\mathbf{x} + \mathcal{S}_\epsilon$. PGD conducts a fine-grained perturbation on images and can achieve a higher attack success rate than FGSM under the maximum perturbation magnitude (i.e., the perturbation budget), at the cost of a longer running time.

The perturbation budget [22], [91] is a pivotal concept in adversarial attacks. It signifies the maximum permissible alterations that an adversary is allowed to implement on the input data. This constrains the extent of the perturbations that the adversary can apply, thus maintaining the attack within realistic and manageable bounds. Specifically, we consider a (potentially stochastic) mapping $\rho : \chi \rightarrow \chi'$, where χ and χ' are vector spaces. In an untargeted adversarial attack aiming to induce a misclassification by the classifier *C*, the goal is to have $C(\mathbf{x}+\rho(\mathbf{x})) \neq C(\mathbf{x})$, while maintaining the perturbation constraint $||\rho(\mathbf{x})||_p \leq \epsilon$, where ϵ denotes a perturbation budget [22]. Given that the outcome of a perturbation is a vector, the perturbation is typically converted to a scalar by applying a *p*-norm operation $||\cdot||_p$. The most frequently used values of *p* include 0, 1, 2, and ∞.

To produce an effective adversarial example within a perturbation budget, the attacker can progressively perturb the elements of an example (e.g., the pixels of an image), by running more sophisticated adversarial attack algorithms, e.g., PGD [22]. The attacker repeats this process until the DNN-based classifier misclassifies the example $\mathbf{x}$ into an attacker-specified target class that differs from the source category of the benign example $\mathbf{x}$.

The information necessary to undertake an adversarial attack against a DNN is relatively easy to obtain. This is due to the fact that the best-performing DNN-based classifiers generally use well-known architectures, e.g., ResNet models [92], and commonly employ the Cross-Entropy loss function [93] for classification tasks. Even in the case where the parameters of the DNN-based classifiers are not accessible for the attacker, it is possible for the attacker to learn a good surrogate of the DNN-based classifiers by sending queries to the classifiers and collecting responses [94].

The attacker can also use other strategies, such as constrained optimization-based or heuristic approaches (see Section III), to seek out effective adversarial instances. After the attacker confirms the effectiveness of the generated adversarial example in a controlled environment, they can launch an actual adversarial attack by feeding perturbed examples to the DNN model under attack.

C. Notable Attack Incidents

In the past several years, there have been several notable adversarial example attacks:

- In 2023, researchers proposed a novel adversarial attack framework [68], and designed to generate adversarial malicious traffic capable of deceiving ML-based traffic classification systems. Experiments demonstrated that this approach exhibits a high evasion growth rate across multiple models and datasets.

- In 2021, researchers demonstrated that adversarial attacks could disrupt DNN-based power distribution in large-scale MIMO network downlinks, and experiments indicated that white-box attacks could result in up to 86% of unworkable solutions [67].
- In 2021 and 2022, researchers demonstrated that they could cause an autonomous vehicle to mistake a stop traffic sign for a speed limit sign by putting a small, almost imperceptible sticker on the sign [95], [96]. While this type of attack has not yet resulted in any real-world accidents, it has raised concerns about the safety of autonomous vehicles and the potential for malicious actors to cause accidents or other harm using adversarial attacks.
- In 2022, a team of researchers showed that adversarial examples could trick image classification systems in self-driving cars [97]. The researchers showed that a small perturbation added to a traffic sign could cause the self-driving car to misidentify the sign, potentially leading to dangerous mistakes on the road.
- In 2019, researchers demonstrated that they could cause an ML model to misclassify a fraudulent credit card transaction as legitimate by applying weak perturbations to the transaction data [98], [99], [100]. This type of attack could potentially result in significant financial losses for financial institutions and consumers.
- In 2020, researchers demonstrated that they could cause a chatbot to generate inappropriate or offensive responses by adding small perturbations to the input text [101]. This type of attack could potentially cause damage to a company's reputation or lead to lost customers.
- In 2022, researchers demonstrated that adversarial examples could trick a voice assistant, e.g., Amazon Alexa and Google Assistant [102]. They manipulated voice commands to make them sound normal to humans but caused voice assistants to perform actions that were not intended. This can sabotage the security and privacy of users who use voices to control smart home devices.
- In 2022, a group of researchers demonstrated that adversarial examples could be utilized to evade spam filters, allowing malicious emails to bypass detection [103], [104]. They created adversarial examples of spam emails by adding perturbations to the email content, and caused the spam filter to incorrectly classify the email as non-spam.

While these adversarial attacks on ML or DNN models have not yet caused widespread financial or economic loss, they have sparked worries about the safety and dependability of these systems, and research into approaches for detecting and defending against these attacks is ongoing.

D. Contributions of This Survey

As the applications of ML and artificial intelligence continue to expand into almost all aspects of human life and society, the robustness and security of ML models become increasingly crucial [105], [106], [107]. As a result, adversarial attacks and defenses make up an active and rapidly growing research area. For example, in the Year 2022 alone, over 1,200 research articles were published on adversarial attacks and defenses, documenting many new attack and defense techniques and incidents. In the Year 2021, over 1,000 research articles were published on these topics.[1] Most of these new research outcomes have not been covered by any of the latest literature reviews and surveys due to the fast pace of this active research area of adversarial attacks and defenses. To this end, a timely summary of emerging attacks and new defense techniques is critical to keep the research community and security practitioners well-informed and equipped with the latest knowledge.

This comprehensive survey delves into the cutting-edge advancements in adversarial attack and defense techniques over the past 24 months. We review over 220 research papers published in Q1 journals classified by the Journal Citation Reports (JCR), indexed by IEEE, ACM, Springer, and Elsevier's ScienceDirect. We also consider papers presented at top-tier conferences, such as AAAI, CCS, and ICCV, since 2021. Our primary focus is on adversarial attacks and defenses that target ML or DNN-based models used in the areas of communication and networking. This survey is anticipated to inspire a new wave of research and innovation in the rapidly evolving field of adversarial attack and defense.

These are the key findings of this survey:

- A comprehensive classification of recent adversarial attack methods as well as the SOTA adversarial defense techniques based on a variety of attack principles, presented in a visually appealing table and tree diagram format.
- The categorization of the methods into counter-attack detection and robustness enhancement, with a specific focus on regularization-based methods for enhancing robustness, represented through tables and tree diagrams.
- A rigorous evaluation of the existing works, including an analysis of their strengths and limitations, and recommendations for future research avenues.

Some of the noteworthy highlights from the survey include:

- An exploration of new avenues of attack in the last two years, including search-based attacks, decision-based attacks, drop-based attacks, and beyond the traditional optimization-based and gradient-based attacks.
- The emergence of physical-world adversarial attacks, particularly in the form of adversarial patches.
- A hierarchical classification of the latest defense methods, highlighting the challenges of balancing training costs with performance, maintaining clean accuracy, overcoming the effect of gradient masking[2] (or in other words, a defense method appears to work but is actually ineffective), and ensuring method transferability.

As illustrated in Fig. 3, this survey is organized as follows. Section II provides a brief overview of the existing

[1] These search results are based on IEEE Xplore with the keyword "adversarial" and "attack" as of 7 February 2023.

[2] Gradient masking here refers to the phenomenon that the gradient of the model is hidden or obsolete, e.g., towards potential adversaries. On the other hand, it also refers to a category of defense techniques that exploit or aim to achieve the phenomenon of gradient masking [108].

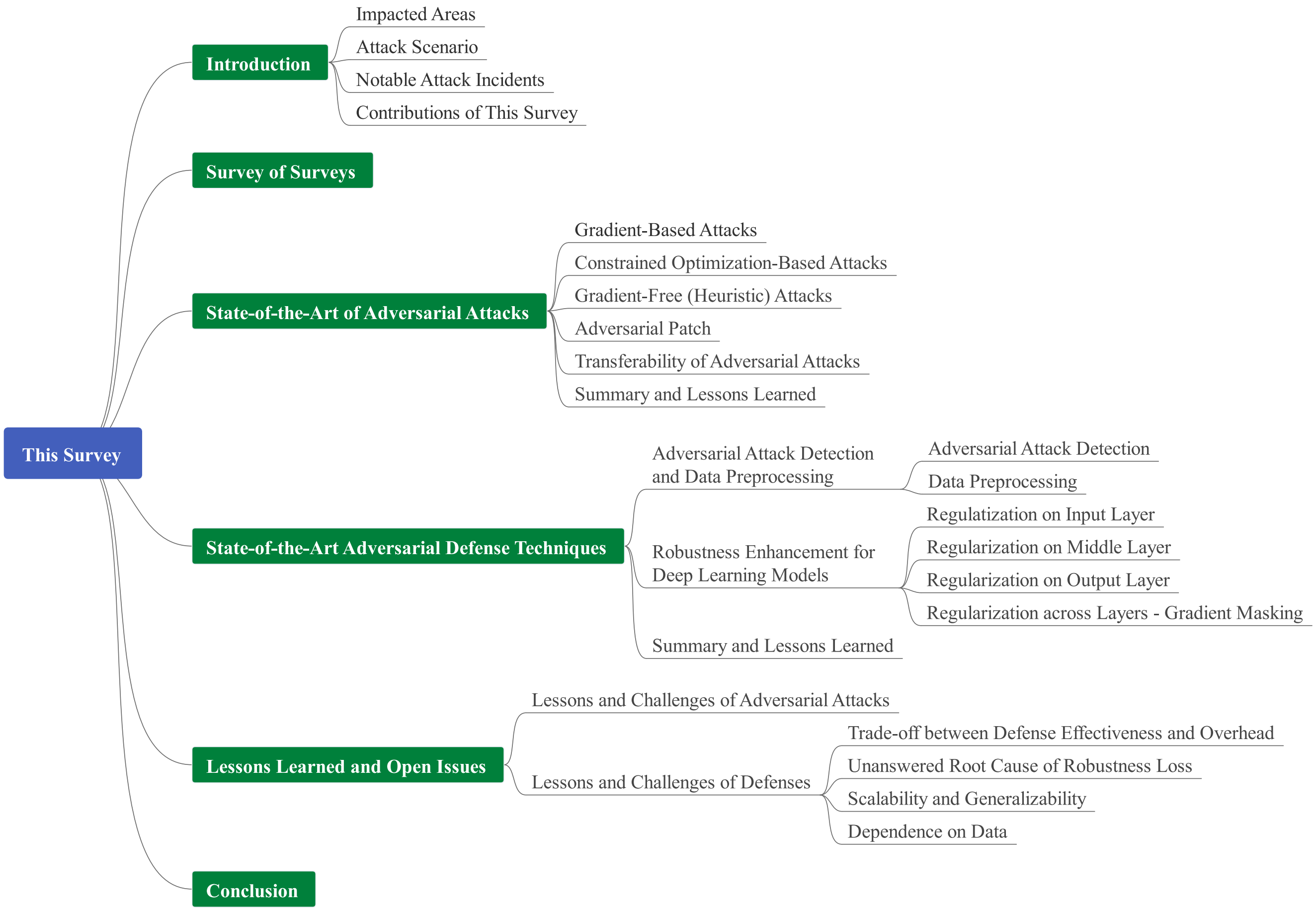

This Survey
Introduction
Impacted Areas
Attack Scenario
Notable Attack Incidents
Contributions of This Survey
Survey of Surveys
State-of-the-Art of Adversarial Attacks
Gradient-Based Attacks
Constrained Optimization-Based Attacks
Gradient-Free (Heuristic) Attacks
Adversarial Patch
Transferability of Adversarial Attacks
Summary and Lessons Learned
State-of-the-Art Adversarial Defense Techniques
Adversarial Attack Detection and Data Preprocessing
Adversarial Attack Detection
Data Preprocessing
Robustness Enhancement for Deep Learning Models
Regulatization on Input Layer
Regularization on Middle Layer
Regularization on Output Layer
Regularization across Layers - Gradient Masking
Summary and Lessons Learned
Lessons Learned and Open Issues
Lessons and Challenges of Adversarial Attacks
Lessons and Challenges of Defenses
Trade-off between Defense Effectiveness and Overhead
Unanswered Root Cause of Robustness Loss
Scalability and Generalizability
Dependence on Data
Conclusion

Fig. 3. The anatomy of this survey.

surveys of adversarial attacks and defenses, and clarifies the key differentiating factors of the current survey. Section III categorizes the most recent adversarial attacks, and provides a comprehensive analysis of each of the categories, as well as the transferability of adversarial attacks. Section IV classifies and analyzes various adversarial defense and detection techniques, and their effectiveness and limitations against the latest adversarial attacks. Lessons learned and open challenges are delineated in Section V. At last, Section VI summarizes the current state and suggests avenues for future investigations. Table I defines the notation used in this survey.

II. Survey of Surveys

Adversarial attacks and defenses in ML and DNN models are crucial areas of research that have garnered significant attention in recent years. There are several reviews on the topic, each delving into specific aspects of the topic. Addressing the profound concern posed by malware in the context of network security, Yan et al. [82] delve into the domain of adversarial attacks and defenses for malware classification utilizing ML techniques. The authors provide a cohesive overview of a unified framework for malware classification, and present an exhaustive examination of ML-based approaches for malware classification, encompassing adversarial attacks against malware classifiers and robust malware classification. The authors of [110] first present preliminary knowledge concerning adversarial examples, and then contrast theoretical models of adversarial example attacks with actual instances of attacks. They also present existing examples of actual adversarial attacks. The authors of [109] review adversarial attacks and defenses in the computer vision domain, as well as their real-world applications. They analyze the various methods proposed for attacking and defending against adversarial attacks in this domain and explore these methods' effectiveness and limitations. Similarly, the authors of [112] discuss the theoretical underpinnings, methods, and applications of adversarial attack techniques. In addition, they present several research initiatives on defensive strategies that span a broad variety of frontiers in the area, followed by a discussion of a number of open issues and challenges. The authors of [113] expand the scope of their review to include adversarial attacks in the context of images, malicious code, and text across various domains. They discuss the various types of adversarial attacks proposed in these contexts and analyze their effectiveness. The authors of [111] focus on summarizing the recent studies on adversarial attack and defense techniques in the deep learning area. They study existing defense methods from three perspectives: Data altercation, model modification, and utilization of auxiliary tools. They analyze the benefits and drawbacks of each strategy and discuss the limitations of existing methods.

TABLE I
NOTATION AND DEFINITION

Notation	Definition
$\mathbf{x}$, y	A clean input example and its ground-truth label.
y_t	The label of the class designated by an attacker.
$\mathbf{x}^{adv}$, t	An adversarially perturbed example and the attacker-specified target class.
$\mathbf{x}^{(i)}$	An adversarial example generated in i-th round of processing.
$\tilde{\mathbf{x}}$	The optimal state in the Markov decision process tackled by reinforcement learning.
$\epsilon \in \mathbb{R}^+$	The magnitude of perturbation applied to an example.
$\nabla_{\mathbf{x}}\mathcal{L}(\cdot,\cdot)$	The gradient of the loss function $\mathcal{L}(\cdot,\cdot)$ against $\mathbf{x}$.
α, $\alpha^{(i)}$	A constant step size used for iterative adversarial example generation, and a variable step size used in the i-th iteration of generation, respectively.
ϵ	The maximum perturbation magnitude, i.e., the perturbation budget.
$\mathcal{S}_\epsilon \subseteq \mathbb{R}^d$	The set of allowed perturbations under the maximum perturbation magnitude ϵ.
$\prod_{\mathbf{x}+\mathcal{S}_\epsilon}(\cdot)$	The projector that maps its input to the closest element to the input in the set $\mathbf{x}+\mathcal{S}_\epsilon$.
$\lVert \cdot \rVert_p$, ℓ_p	p-norm, where p can be 0, 1, 2 or ∞.
g_i	The gathered gradient in the i-th iteration.
z_i	The logit of example class i
$f(\cdot)$	The predict function built from an ML or DNN model.
$\text{Clip}_{\mathbf{s}}(\cdot)$	The operation that clamps the elements of the input to within $\mathbf{s}$.
ξ_i^σ	The random variables drawn i.i.d. from a distribution P^σ parameterized by the standard deviation $\sigma \in \mathbb{R}^+$.
η	The neighboring hypothesis in IoU.
$\mathcal{L}^{sce}$	The softmax cross-entropy (SCE) discrepancy between the one-hot ground truth and the output.
$\mathcal{L}_{T_i}(\cdot)$	The loss function of a reinforcement learning task T_i.
I	The maximum number of gradient-update iterations.
$\mathbf{z}^{(l)}$	The feature derived from the l_{th} layer.
$\bigtriangledown$	The gradient calculation.
$\odot$	The Hadamard product.
$S \subset \mathbb{R}^n$	A subset of n-dimensional real number space.

There are also many investigations focusing on real-world attacks. In [118] and [117], the authors focus on physical adversarial attacks. They classify and summarize current physical adversarial attacks from the perspective of physical world attacks, discussing the benefits and limitations of various approaches. In [40], the authors examine adversarial attacks and defenses against transaction records from the viewpoint of NLP.

Other existing surveys are concerned with techniques for enhancing the robustness and resilience of DNN models in the face of adversarial attacks. For example, in [115], the authors analyze and compare adversarial training methods. They discuss the various approaches proposed for adversarial training and analyze their effectiveness in enhancing the robustness of deep learning models.

These above-mentioned earlier reviews, e.g., [109], [110], focus more on classical attack and defense approaches. Conversely, with the advancement of deep learning in the last two years, more and more new risks have emerged. For example, Gallagher et al. [119] adapt FGSM as a single value and label flipping attack on financial stock data-based prediction networks, and find that it can result in a significant reduction in profitability and financial losses in a financial trading simulation. It is highlighted that the potential consequences of manipulating stock prices through buying and selling in the public trading market could be significant. Goldblum et al. [120] examine the impact of adversarial attacks on trading robot-based stock price predictions. These systems, known as high-frequency trading (HFT) systems, operate in extremely short time frames, making it difficult to prevent harmful behavior through human intervention. This is particularly concerning as it is well accepted that irregular behavior and security breaches in HFT systems have precipitated major market incidents like "Flash Crash" [121]. The severity of the damage caused by adversarial attacks in such systems cannot be underestimated.

As summarized in Table II, this survey aims to bridge the gaps in the current literature by not only focusing on classical methodological analysis but also systematically examining new methods that have emerged in the last two years. Further, Table III illustrates the gap between the existing surveys on adversarial attacks and this survey. Specifically, this survey reviews the new methods in the context of adversarial attack methods, classifying them in a new light, and documenting effective but under-reported new attack methods, such as decision-based and drop-based methods. It also sheds light on the recent developments in adversarial patching, a powerful physical world attack that has been under-explored in previous surveys. From a defensive perspective, the survey covers both adversarial detection methods and model robustness enhancement methods, categorizing them from a novel perspective starting from the hierarchy of operations, summarizing them in the form of tables and tree diagrams, and suggesting future research directions. The survey provides an in-depth understanding of the SOTA in adversarial attacks and defenses in deep learning.

III. State-of-the-Art of Adversarial Attacks

An adversarial attack is a deliberate intent to mislead an ML or DNN model by introducing subtle, imperceptible interference to an input sample. This might result in the model drawing an incorrect conclusion confidently. Szegedy et al. [75] were among the first to discover that DNNs are susceptible to slight adversarial perturbations. Ever since, considerable efforts have been committed to producing more potent adversarial attacks for evaluating the robustness of DNNs.

Conventionally adversarial attacks consist of black-box attacks [108], white-box attacks [122], and gray-box attacks [123]. A black-box attack signifies that an attacker has no knowledge of the underlying structure, learnable parameters, or defense strategies of the model under attack. The attacker interacts only with the model via its inputs and outputs [108]. A white-box attack occurs when the attacker has all prior knowledge of the model under attack, e.g., the loss function and the optimized parameters of the model, and exploits the knowledge to facilitate the attack [122]. A gray-box attack accounts for the case, where the attacker only possesses partial knowledge of the model under attack in prior [123].

In this section, we classify the existing adversarial attacks from their underlying mechanisms and implementation techniques. Some of the attacks can be adapted to support some

TABLE II
COMPARISON OF THE EXISTING SURVEYS ON ADVERSARIAL ATTACKS AND DEFENSES

Reference	**Year**	**Focus Area**					
		Digital-world Attacks	**Physical-world Attacks**	**Defenses**	**Detection**	**Transferability**	**No. of references since 2021**
[109]	2018	✓	✓	✓	✓	✗	n.a.
[110]	2018	✓	✓	✗	✗	✗	n.a.
[111]	2019	✓	✗	✓	✓	✓	n.a.
[112]	2020	✓	✓	✓	✓	✗	n.a.
[40]	2021	✓	✓	✓	✗	✗	n.a.
[113]	2021	✓	✓	✓	✓	✗	4
[114]	2022	✓	✗	✓	✓	✗	1
[115]	2022	✓	✗	✓	✗	✗	5
[116]	2022	✓	✓	✓	✓	✗	16
[117]	2022	✗	✓	✓	✓	✓	38
[118]	2022	✗	✓	✗	✗	✓	42
[82]	2023	✓	✗	✓	✓	✗	17
This survey	2023	✓	✓	✓	✓	✓	221

TABLE III
THE GAP ANALYSIS OF THE EXISTING SURVEYS ON ADVERSARIAL ATTACKS AND DEFENSES. WHILE THIS SURVEY FOCUSES ON COMMUNICATIONS AND NETWORKING, IT ALSO HAS SIGNIFICANT COVERAGE OF ADVERSARIAL ATTACKS AND DEFENSES DESIGNED FOR OTHER APPLICATIONS (E.G., COMPUTER VISION OR NLP) BUT MAY APPLY TO COMMUNICATIONS AND NETWORKING

Surveys	Communication & Network			Computer Vision			Natural Language Processing		
	2021	2022	2023	2021	2022	2023	2021	2022	2023
ours	16	17	17	91	61	15	3	8	0
[82]	15	2	0	38	12	0	0	0	0
[113]	0	0	0	4	0	0	0	0	0
[114]	0	0	0	1	0	0	0	0	0
[115]	0	0	0	5	0	0	0	0	0
[116]	0	0	0	16	0	0	0	0	0
[117]	0	0	0	38	12	0	0	0	0
[118]	0	0	0	24	17	0	1	0	0
[109]	0	0	0	0	0	0	0	0	0
[110]	0	0	0	0	0	0	0	0	0
[111]	0	0	0	0	0	0	0	0	0
[112]	0	0	0	0	0	0	0	0	0
[40]	0	0	0	0	0	0	0	0	0

or all of the black-, white-, and gray-box attacks, as briefly described in the following.

- Gradient-Based Attacks: Gradient-based attacks manipulate the input data to ML or DNN models based on the gradient of the model's loss function. This leads to an increase in the loss function, thereby causing the model to make erroneous predictions. Since attackers need to access the gradient, these attacks are more commonly associated with white-box attack scenarios. For instance, an attacker may download the target model after compromising the server on which the model is running. See Section III-A.
- Constrained Optimization-Based Attacks: These attacks model the generation of adversarial samples as an optimization problem with a perturbation magnitude constraint. They are more likely to occur in a gray-box attack setup, since the optimization process can only use partial information about the target model. For instance, attackers might obtain the architecture of the target model through information gathering or best-practice analysis. See Section III-B.
- Gradient-Free (Heuristic) Attacks: These attacks do not utilize mathematical models. Instead, they use human intuition and expert knowledge to obtain adversarial samples that can deceive the target ML or DNN models. They are typically employed in a black-box setup and can be further classified into search-based, decision-based, and score-based methods. See Section III-C.
- Adversarial Patch: Adversarial Patch abandons the constraint of imperceptible perturbation to benign samples. In the meantime, it manages to make the adversarial change appear normal, deceiving both humans and machines. An Adversarial Patch can be constructed in white-box, gray-box, or black-box settings. See Section III-D.

A. Gradient-Based Attacks

Gradient-based attacks are a common type of attack used against neural network models. These attack methods work by manipulating the input data according to the gradient of the loss function regarding the input to cause the model's loss function to increase, effectively causing the model to make

TABLE IV
BRIEF SUMMARY OF EXISTING SURVEYS ON ADVERSARIAL ATTACKS

Survey	Year	Focus Area
This survey	2023	• A comprehensive classification of recent adversarial attack methods and the SOTA adversarial defense techniques based on various attack principles presented in a visually appealing table and tree diagram format. • Categorization of the methods into adversarial attack detection and robustness enhancement, with a specific focus on regularization-based methods for enhancing robustness, represented through tables and tree diagrams. • A rigorous evaluation of the existing works, including an analysis of their strengths and limitations, and recommendations for future research avenues.
[82]	2023	• Presents a framework which summarizes the five general phases of malware classification. • Systematic survey on malware classification, adversarial attacks, and robust malware classification, highlighting the evolution of adversarial attack and defense in the Defense-Attack-Enhanced-Defense process.
[115]	2022	• Introduces robust adversarial training to defend against adversarial samples. • Connections to traditional machine learning theories are investigated. • Different approaches with adversarial attacks and defense/training algorithms are summarised. • Presents analysis, outlook and comments on adversarial training.
[118]	2022	• Examines the evolution of physical adversarial attacks in computer vision applications using DNNs, e.g., image recognition, object detection, and semantic segmentation. • The survey classifies the present physical adversarial attacks, focusing on strategies employed to keep adversarial features in physical contexts.
[117]	2022	• Classifies physical attacks in terms of attack task, attack form, and attack method. • Classifies adversarial defenses in terms of pre-processing, intra-processing and post-processing of DNN models. • Challenges of this research area are discussed, and present some future research directions.
[114]	2022	• Overviews the fundamentals and features of adversarial attacks and evaluates recent adversarial instance-producing mechanisms. • In-depth discussion of adversarial instance defense mechanisms from three aspects: model, data, and network. • Challenges and outlooks in the field are presented within the context of the present development of adversarial instance generation and defense techniques.
[116]	2022	• A comprehensive survey of the latest advancements in the adversarial robustness of models for object detection is provided. • Prominent attack and defense approaches are reviewed, their advantages and disadvantages are discussed, • A review of recent literature on adversarial robustness in the context of autonomous vehicles is conducted.
[113]	2021	• Illustrates the importance of adversarial attacks, and outlines the key ideas, categories, and dangers of adversarial attacks. • Typical attack and defense techniques for various application areas are reviewed. • Focuses on images, text and malicious codes, presents open questions, and conducts a comparison study with other relevant surveys.
[40]	2021	• Adversarial attacks and defenses against transaction record data are investigated. • Analyses the distinct structure of transaction records compared to typical NLP or time series data and summarizes the characteristics of transaction records and their impact on adversarial attacks. • A scenario for black-box attacks is considered, focusing specifically on adding transaction tokens to the sequence's end.
[112]	2020	• Presents the theoretical basis, algorithms, and practical uses of adversarial attack algorithms. • Considerable research efforts on defensive techniques are described, which cover a wide range of frontiers within the arena. • Several challenges and open issues are discussed.
[111]	2019	• Explains adversarial attack approaches in both the training and testing phases. • Adversarial attack methods are sorted out by their applications in computer vision, NLP, cyber security, and the real world. • Adversarial defense methods are organized into three categories: Data modification, model modification, and the use of auxiliary tools.
[110]	2018	• Provides some fundamental information about adversarial examples. • The theoretical model for adversarial example attacks is contrasted with the real-world model. • Existing practical examples of adversarial attacks are presented.
[109]	2018	• Works on designing adversarial attacks are reviewed, and the occurrence of the attacks is analyzed. • Defenses against these attacks are proposed. • Separately reviews the contribution of evaluating adversarial attacks in real-world scenarios. • Offers a more comprehensive perspective on this field of research.

errors in its predictions. Classic gradient-based attack methods include FGSM [54], Basic Iterative Method (BIM) [124], PGD [22], JSMA [73], DeepFool [76] and others. Numerous recent endeavors have been made to design and create new gradient-based attacks. Table V categorizes the latest gradient-based attacks from the perspectives of input, attack type, and invisibility metric (or "inv-metric" for brevity).

The evolution of gradient-based adversarial attacks has significantly infiltrated various domains of communication networks. A black-box gradient estimation method for network intrusion detection models is natural evolution strategies (NES) [5], where the estimated gradient can be used to project gradient descent (as used in white-box attacks) to build adversarial examples. This approach does not require a proxy network, so queries are more efficient and reliable when crafting adversarial examples. By extending FGSM, momentum iterative FGSM, and projected gradient descent adversarial attacks in FGSM systems, Manoj et al. [67] demonstrate that adversarial attacks can disrupt ML-based power distribution in massive MIMO network downlinks. A recent study in [1] scrutinizes the vulnerability of DNN-based modulation recognition models within communication systems subject to adversarial attacks. Their meticulous examination involves the construction of high-precision DNN-based models, executing multiple adversarial attacks on well-trained models, and unveiling the substantial threat that adversarial

TABLE V
GRADIENT-BASED ADVERSARIAL ATTACK METHODS

Attack	Short Description	Input	Attack type	Invisibility Metric	Advantage	Disadvantage
IGA [133]	A new attack method for link prediction that utilizes gradient information from a trained GAE model.	Discrete	White-box	cross-entropy	IGA achieves effective attack results regardless of whether the global graph information is complete or not. Strong transferability on different realistic diagrams.	The algorithmic complexity dramatically increases when the size of graphs grows larger.
SGA [134]	A framework that reduces the scale of the graph to a smaller subgraph centered around the target node, resolving the challenge of performing attacks on large graphs.	Discrete	Black-box	DAC	Significantly enhances time and memory efficiency and attaches considerable attack strength; Strong transferability among different commonly used graph neural networks.	Not available for node injection attack due to the costly computation of DAC; Often used for node classification or targeted attacks.
Contrastive AT [139]	A generative scheme named Contrastive Adversarial Training is designed, inspired by HMCAM, which aims to produce a series of adversarial examples in a single run.	Continuous	White-box, Black-box	ℓ_∞	Its ASR is superior to other black-box models, equivalent to other white-box models, strikes a compromise between efficiency and accuracy in AT, and demonstrates enhanced transferability.	In the ImageNet, the ResNet-50 model trained with the PGD algorithm takes one-third of its time, but the best robust accuracy is inferior.
IPW&IBA [126]	Creates a cost function that penalizes a subset of feature activations. To determine the real gradient of the black-box target model, computes the directional derivatives along the non-redundant previous directions using an iterative zeroth-order optimization procedure.	Continuous	White-box, Black-box	KL, PLC, NSS, ℓ_1	The best trade-off between attack capabilities and redundancy compared to other white-box attacks, making a good trade-off between black-box attack capabilities and query costs.	While speeding up black-box attacks, it currently faces challenges in meeting the demands of time-sensitive applications, such as AT, which requires a substantial amount of adversarial examples.
DSNGD [127]	Computes the weighted mean of previous gradients from the optimization history to determine the gradient direction of an adversarial attack.	Continuous	White-box, Black-box	ℓ_∞	The sampling operation's computing overhead is reduced, resulting in greater efficiency. Less vulnerable to noise and local optimization, and more accurately approximate the global upward direction.	The performance of the method on larger datasets, e.g., ImageNet, has not been determined.
AoA [129]	Based on the semantic properties shared by DNN. Unlike other methods that focus on attack output, AoA changes the attention heat map and the loss function.	Continuous	White-box, Black-box	RMSE	Beats many DNNs with zero queries. Increased transferability when using traditional cross-entropy loss instead of attention loss. Easy to combine with other transferability enhancement technologies to achieve SOTA performance.	Even though the generated examples are distinct from others, they can still be captured by adversarial detection.
LAFEAT [125]	LAFEAT algorithm takes advantage of latent features in its gradient descent steps.	Continuous	White-box	ℓ_∞	Seeks to harness latent features in a generalized framework. Computationally efficient.	It remains unclear how latent features can be leveraged as viable attack vectors.
SCA-based [141]	A gray box attack method using SCA to predict model structures based on pre-trained classifiers.	Continuous	Gray-box	ℓ_p	The decision boundary of a trained gray-box alternative model is nearer to the target model. More effective than a black-box attack, and more practical compared to a white-box attack.	For complex architectures, the algorithm can be time-consuming and resource-intensive.
SRLIM [138]	An approach that uses SRLIM to preserve the topology in proxy embedding and thereby improves the performance of a gradient-based attacker in a non-target poison gray-box scenario for adversarial attacks.	Discrete	Gray-box	ℓ_0	SRLIM enables the proxy model to learn topologies through isometric mapping, thereby improving the reliability of gradients utilized in the attack models and the transferability.	As the complexity of the graph model structure rises, the computational complexity grows exponentially.
AtkSE [142]	An attack model that integrates semantic invariance modules and momentum gradient ensemble modules to reduce errors within the structural gradients	Discrete	Gray-box	ℓ_0	The gradient fluctuation in semantic graph enhancement and the instability of proxy models are addressed. It improves the attack intensity of the attacker and ensures the transferability of the gray-box attack.	The trade-off between computational efficiency and error reduction is also worth further study.

samples can pose to DNN-based modulation recognition models.

Yu et al. [125] discover that certain “robust” models have hidden features that are unexpectedly susceptible to adversarial attacks. They propose latent feature attack (LAFEAT), a unified ℓ_∞-norm white-box attack algorithm that uses latent features throughout gradient descent steps for computationally efficient attacks, which can be cast as

$$\max_{h,\lambda,\alpha,\mathcal{L}^{sur}} \mathcal{L}^{sce}\Big(f_\theta\Big(\mathrm{PGD}_{\epsilon,x,y}\Big(\mathcal{L}^{lf}_\lambda, \alpha, I\Big)\Big), y\Big),$$
$$where\ \ \mathcal{L}^{lf}_\lambda(\mathbf{z}) = \mathcal{L}^{sur}\left(\sum_{l=1}^{N} \lambda^{(l)} h^{(l)}\Big(\mathbf{z}^{(l)}\Big), y\right). \tag{2}$$

Here, $\mathcal{L}^{sce}$ represents the softmax cross-entropy (CE) loss between the one-hot truth value y and the output. The constant I defines the maximum number of iterations of gradient update iteration. For each layer $l \in \{1, \ldots, N\}$, the value $\lambda^{(l)} \in [0, 1]$ is assigned to the gradient of the layer, and the sum of all values is equal to 1. $\mathbf{z}^{(l)} = f^{(l)} \circ \cdots \circ f^{(1)}(\mathbf{z})$ indicates the feature obtained from the l_{th} layer. The mapping $h^{(l)}$ maps the features from $f^{(l)}$ to the logits for the l-th layer. For ease of exposition, Yu et al. [125] define

$$\mathrm{PGD}_{\epsilon,\mathbf{x},\mathbf{y}}(\mathcal{L}, \alpha, i) = \hat{\mathbf{x}}_i, \tag{3}$$

where $\hat{\mathbf{x}}_i$ is obtained by running the PGD algorithm [22]. PGD identifies an adversarial instance by iteratively updating:

$$\hat{\mathbf{x}}_{i+1} = \mathcal{P}_{\epsilon,\mathbf{x}}\big(\hat{\mathbf{x}}_i + \alpha_i\, sign\big(\nabla_{\hat{\mathbf{x}}_i} \mathcal{L}^{sce}(f_\theta(\hat{\mathbf{x}}_i), \mathbf{y})\big)\big). \tag{4}$$

Here, $\mathcal{I} \in \mathbb{R}^{C\times H\times W}$ limits the image data to a valid range, the function $\mathcal{P}_{\epsilon,\mathbf{x}} : \mathbb{R}^{C\times H\times W} \longrightarrow \mathcal{I}$ clips its input into the ϵ-ball neighbor that is denoted as $\mathcal{I}$. The term $\nabla_{\hat{\mathbf{x}}_i} \mathcal{L}^{sce}(f_\theta(\hat{\mathbf{x}}_i), \mathbf{y})$ calculates the loss' gradient regarding the input $\hat{\mathbf{x}}_i$. α_i indicates the step size. At last, $sign(\cdot)$ is a sign function and returns "–1", "0", or "1" for each element of the gradient.

The objective of LAFEAT is to determine the optimal combination of logit mappings $h = (h^{(1)}, \ldots, h^{(N)})$, their respective weights $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(N)})$, the size of the steps in the schedule α, and the choice of the surrogate loss $\mathcal{L}^{sur}$ to use. Nonetheless, the efficient utilization of latent features as novel attack vectors has not yet been completely comprehended.

In the context of modulation classification, Zhang et al. [2] gauge the performance of Transformer-based neural networks in terms of classification, as well as their susceptibility and resilience to adversarial attacks. Utilizing real datasets, they demonstrate the superior accuracy of Transformers over CNNs when confronted with adversarial attacks. Considering DNN-based modulation classification, Manoj et al. [63] introduce random smoothing, hybrid projection gradient descent adversarial training, and fast adversarial training to create DNN models that are robust to attacks and evaluate them under white-box and black-box attacks. Kotak and Elovici [52] apply adversarial attacks to assess the vulnerability of ML-based IoT device identification systems. The findings in [52] reveal that a novel methodology employing heatmaps to generate adversarial examples can deceive these systems with remarkable effect.

Che et al. [126] propose an Iterative Partially White-box subspace attack (IPW). This technique establishes the cost function in the key hidden space, where the receptive field is at its peak. The cost function penalizes the part of the feature activations that corresponds to both salient and guiding regions, rather than penalizing each pixel across the comprehensive dense output space. Moreover, they present an Iterative Black-box attack (IBA). This approach employs non-redundant variations from original models as initial hints to gauge the gradient of a target black-box model. The estimation is done through a zeroth-order iterative optimization process that computes the directional derivatives along the initial directions that are not redundant.

Fig. 4 offers the overview of the IPW&IPA framework developed in [126]. The first step illustrates the concept of a subspace assault by producing a non-repetitive initial perturbation from a partial white-box source model. To deceive a target model that is unknown to them, they merge an *a-priori* optimizer with a zero-order optimizer in Step 2. The method balances the attack capability and perturbation redundancy, overcoming the issues of costly attack cost and imperceptibility. Although the proposed non-repetitive initial cues enhance the black-box attacks, it remains challenging to fulfill the demands of certain time-sensitive applications, e.g., adversarial training necessitating a large quantity of adversarial instances.

Dynamically Sampled Nonlocal Gradient Descent (DSNGD) [127] computes the gradient direction for an adversarial attack by calculating the weighted mean of previous gradients from an optimization record. The gradient computation in DSNGD can be written as

$$\nabla_x \mathcal{L}(f_\theta(x_t), y) := \sum_{i=1}^{t} w_i \cdot \nabla_x \mathcal{L}(f_\theta(\hat{x}_i), y),$$
$$\hat{x}_i = \mathrm{Clip}_{[0,1]}(x_i + \xi_i^\sigma). \tag{5}$$

Here, $\mathcal{L}$ indicates the loss function, such as CE, of a neural network f_θ. $\mathrm{Clip}_{[0,1]}(\cdot)$ clamps the input to the range between 0 and 1; $\hat{x}_i$ denotes a noisy sample in the optimization process; w_i refers to the gradient weight associated to $\hat{x}_i$; the random variables ξ_i^σ are taken from the i.i.d. distribution P^σ parametrized by the standard deviation $\sigma \in \mathbb{R}^+$. The variable t stands for the iteration number during the current attack. This improves the efficiency of the algorithm by reducing the computational burden caused by sampling operations, eliminating the need for manually tuning additional hyperparameters, and providing a more precise estimation of the overall upward direction. However, its performance on larger datasets, e.g., ImageNet, is yet to be determined.

In the endeavor of modulation and recognition of communication signals using CNN, Yang et al. [128] put forth a white-box attack algorithm known as the shortest distance algorithm (SD-Alg). This innovative approach can generate minimal interference and considerably degrade the CNN model's classification performance. Chen et al. [129] propose an Attack on Attention (AoA) method that depends on the semantic characteristics common among multiple DNNs to enhance the transferability of adversarial attacks. As opposed to prior techniques that concentrate on attacking the output, such as the One-Pixel attack developed in [130], AoA aims to modify the attention heat map and achieves exceptional results in black-box attacks. This method produces adversarial instances that can deceive numerous DNNs using zero queries and leads to a substantial improvement in transferability if the standard CE loss is substituted with an attention loss. The AoA attack can be seamlessly integrated with other

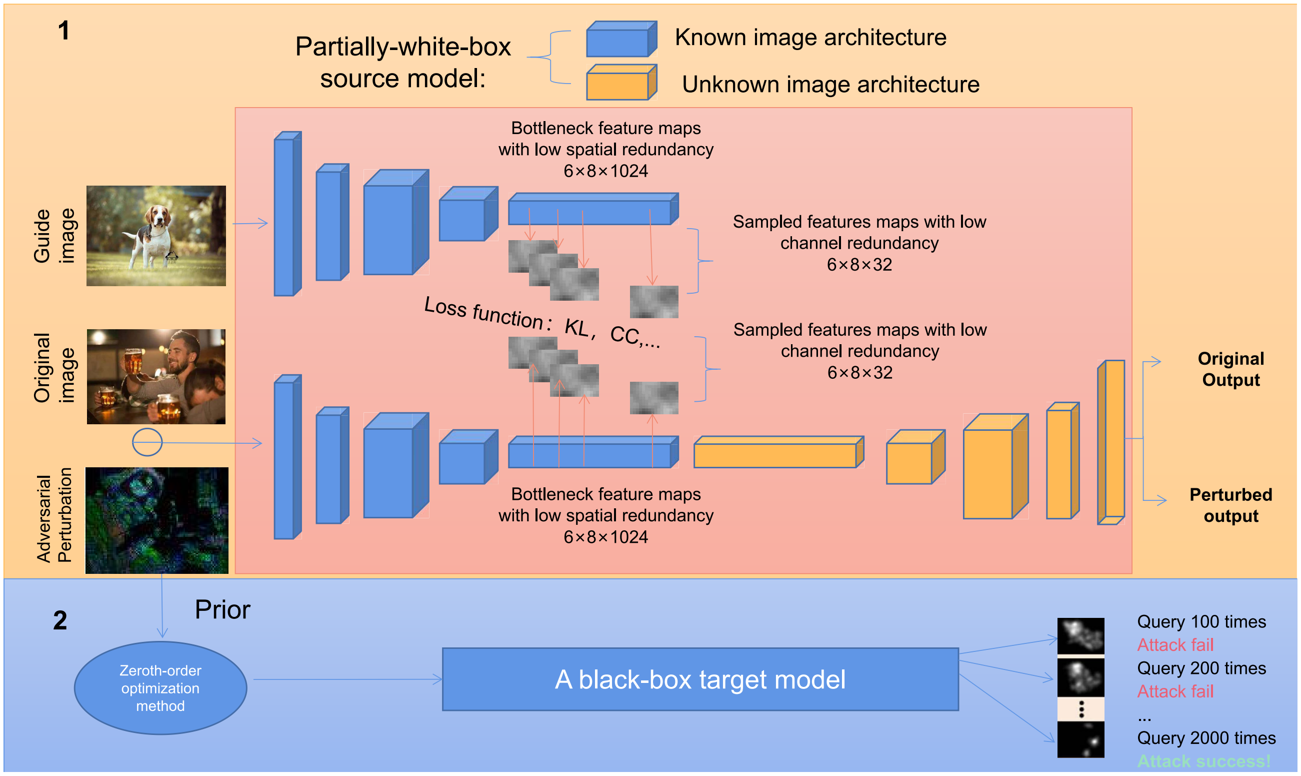

1
Partially-white-box source model:
Known image architecture
Unknown image architecture
Bottleneck feature maps with low spatial redundancy 6×8×1024
Sampled features maps with low channel redundancy 6×8×32
Loss function: KL, CC,...
Sampled features maps with low channel redundancy 6×8×32
Guide image
Original image
Adversarial Perturbation
Original Output
Perturbed output
Bottleneck feature maps with low spatial redundancy 6×8×1024
2
Prior
Zeroth-order optimization method
A black-box target model
Query 100 times
Attack fail
Query 200 times
Attack fail
...
Query 2000 times
Attack success!

Fig. 4. Sketch of the IPW&IBA: The first step illustrates the concept of a subspace assault by creating a non-redundant prior perturbation from a partial white-box source model. To deceive a target model that is unknown to them, they merge an *a-priori* optimizer with a zero-order optimizer in Step 2. [126].

transferability-enhancement methods to attain cutting-edge performance.

Graph structures are common in the physical world, and DNNs are commonly employed to tackle graph network problems, including node classification [131] and link prediction [132]. Iterative Gradient Attack (IGA) [133] is a new approach for link prediction that leverages gradient information from a trained Graph Autoencoder (GAE) model. IGA offers effective results under both complete and incomplete graph information, and it can be integrated with various tasks. IGA also has good transferability across various realistic diagrams. Unfortunately, its computational complexity can grow significantly when the size of the graph increases.

Considering node classification tasks, Li et al. [134] propose a Simplified Gradient-based Attack (SGA), which addresses the difficulty in attacking large-scale graphs by leveraging a subgraph that comprises *k*-hop neighbors of attacked. An input graph is perturbed by sequentially flipping edges whose magnitude of gradients is the biggest in this subgraph. The SGA method overcomes the issue of gradient fading in gradient-based attack techniques by using a smaller subgraph centered around the node under attack and by incorporating a scaling factor. As depicted in Fig. 6, SGA has a significant advantage over other cutting-edge attack techniques, e.g., the GradArgmax method developed in [135] and the Nettack method developed in [136], in terms of time and memory efficiency, meanwhile still achieving considerable attack results.

On the other hand, a novel attack scenario is known as a node injection attack, in which attackers can inject a set of malicious nodes into a graph to circumvent the original graph's topology and misclassify victim nodes [137]. SGA is generally inapplicable to the node injection attack because the injected malicious node is a singleton node and is not initially linked to any nodes, where a *k*-hop subgraph cannot be extracted, and its high computational cost is also a drawback. Currently, SGA is limited to node classification tasks and targeted attack scenarios. Ongoing research is expected to expand the method and make it more adaptable to various graph analysis tasks.

Gradient-based attackers collect gradients of node features and graph structures and produce perturbations based on them using pre-trained Graphic Neural Network (GNN) classifiers, referred to as proxy models. However, the majority of existing work [133], [137]. Concentrate on using gradients to produce perturbations rather than looking at how to get more dependable gradients from different models. The gradient-based perturbations manufactured by the attacker are affected by the proxy model's embedding layer mapping. The perturbations generated are model-specific, and lose their generalization to another model. To solve this problem, Wang et al. [138] propose Surrogate Representation Learning with Isometric Mapping (SRLIM) to enable the model to learn topologies. Keeping the similarity of nodes from the input layer to the embedding layer, SRLIM passes topological knowledge to the embedding layer, thus improving the effectiveness of the adversarial attacks produced by gradient-based attackers in non-target poison gray-box attacks. However, as

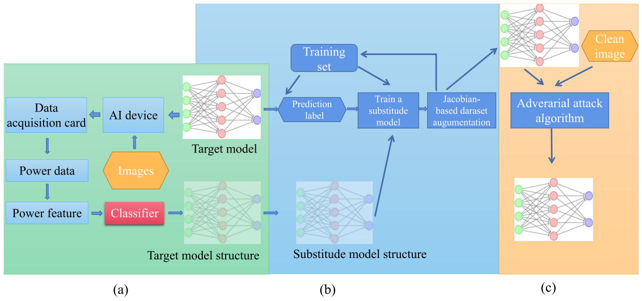


Fig. 5. Working flow chart of the SCA-based attack. (a) Obtain the network structure of the target model. (b) Training an alternative model. (c) Creating an example of hostility.

the complexity of the graph structure rises, the computational complexity will also increase exponentially.

In contrast to the prevalent adversarial attack methods that generate only one adversarial instance for an input, Wang et al. [139] introduce an attack method that produces a range of adversarial examples for a given input. This is achieved through the use of Hamiltonian Monte Carlo with Accumulated Momentum (HMCAM). They also present a novel generative technique, namely Contrastive Adversarial Training (Contrastive AT), which employs Hamiltonian Monte Carlo (HMC) to simulate the creation of adversarial examples and achieves an equilibrium distribution of adversarial instances within just a few rounds by performing some moderate changes of the conventional Contrastive Divergence [140]. As a result, Contrastive AT strikes a balance between attack accuracy and efficiency in adversarial training. The experimental outcomes demonstrate that Contrastive AT attains a higher attack success rate (ASR) than black-box models, and is on par with other white-box models.

Xiang et al. [141] come up with a straightforward and efficient gray-box attack strategy based on the side-channel attack (SCA) policy. The SCA-based attack is illustrated in Fig. 5. First, the target model's fundamental network structure is derived using an SCA-based attack. The alternative model is then trained using the derived network structure. Adversarial samples are produced using the trained alternative parameters in order to mislead the target model. The trained gray-box replacement model's decision boundary is nearer to the target model because gray-box attacks use abundant internal information, as opposed to black-box attacks. It is thus more realistic than a white-box attack and more efficient than a black-box attack. However, there might be more possible architectures in real-world applications. The algorithm must run every step of adversarial and side-channel attacks against every candidate architecture, which can take significant time and resources.

In the face of errors caused by the discreteness of graph structures and subsequent rough gradients, for the vulnerability of GNNs to the semantic space and parameter random initialization resulting in an unstable representation of GNNs, Liu et al. [142] proposed two modules to solve the problem, namely the semantic invariance module and the momentum gradient integration module, and integrated the above modules to propose an attack model named Attacking by Shrinking Errors (AtkSE). This method solves the gradient fluctuation in semantic graph enhancement and the instability of the proxy model to some extent, increases the attack intensity of the attacker, and ensures the transferability of the gray-box attack. But at the same time, the trade-off between computational efficiency and error reduction is also worth further study.

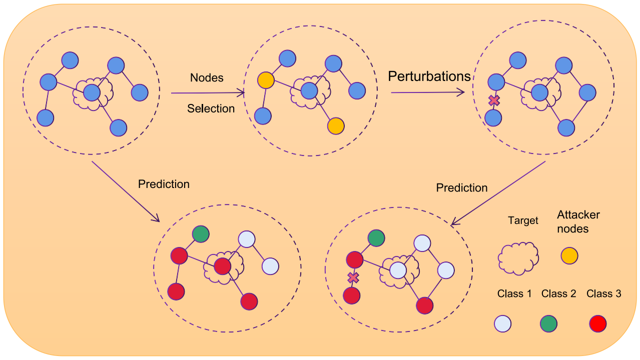


Fig. 6. Iterative gradient-based attack through subgraph expansion. Here, $k = 2$. The dotted circle represents the neighborhood range (k) around the target node, i.e., a subgraph containing the target node. Flipping edges within a subgraph lead to misclassification of the target node.

All of these methods aim to find ways to generate adversarial examples that can fool the DNNs by exploiting their gradients. However, they have different approaches and techniques to leverage the gradients. They also have different strengths and limitations. For instance, some methods take less computational time and memory, such as SGA [134], Contrastive AT [139], and LAFEAT [125], while others are better in terms of transferability and attack performance, such as IGA [133], AoA [129], SRLIM [138], and AtkSE [142].

B. Constrained Optimization-Based Attacks

Attacks based on constrained optimization involve creating adversarial examples by tackling a constrained optimization problem. This method seeks the minimum perturbation, subject to ℓ_0, ℓ_2, or ℓ_∞-norm constraints that cause the neural network model to make an incorrect classification. As such, adversaries can generate an adversarial example $\mathbf{x}^{adv}$ for an untargeted attack (i.e., misclassifying the adversarial example to any different class from the correct one) by following:

$$\begin{aligned} &\max_{\mathbf{x}^{adv}} \ \mathcal{L}\left(f\left(\mathbf{x}^{adv}\right), y\right), \\ &\text{s. t.} \ \left\|\mathbf{x}^{adv} - \mathbf{x}\right\|_\infty \leq \epsilon, \end{aligned} \tag{6}$$

where the objective is to find an adversarial counterpart $\mathbf{x}^{adv}$ and the constraint ϵ specifies the invisibility requirement of adversarial perturbation. $\mathcal{L}(\cdot, \cdot)$ is the loss function of the target model f, and $\mathbf{x}$ is a clean example. In the case of a targeted attack (i.e., misclassifying the adversarial example to an incorrect class designated by the attacker), the objective

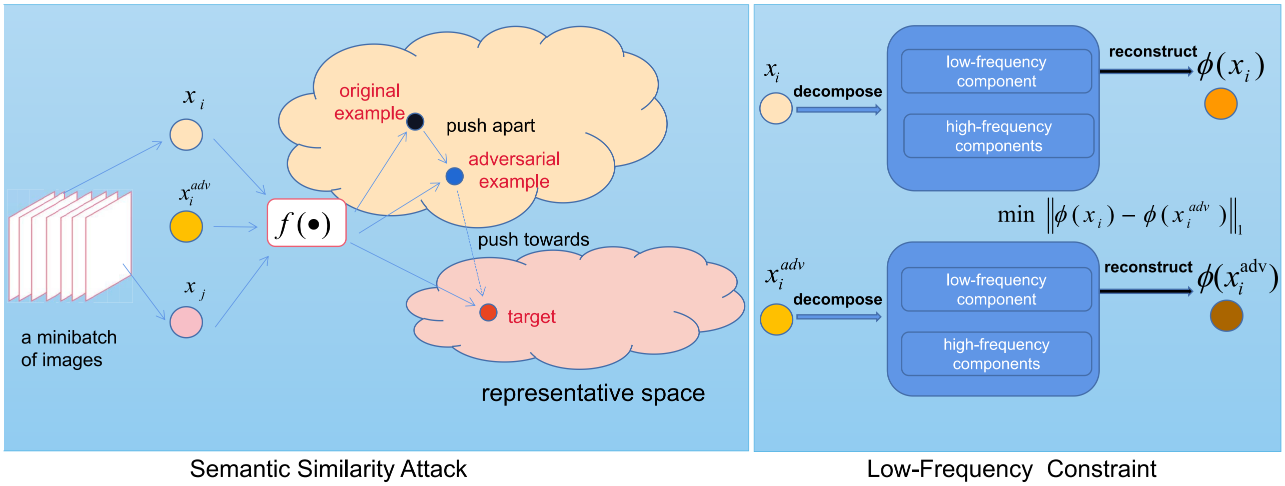


Fig. 7. Overview of the SSAH method. The left subfigure illustrates semantic similarity attacks, while the right subfigure demonstrates the low-frequency constraint. $f(\cdot)$ represents the function mapping from the image to the representation space, and $\phi(\cdot)$ denotes the shallow neural network that divides the image into different frequency components and then reconstructs it by using low-frequency components.

function of the optimization problem is:

$$\min_{\mathbf{x}^{adv}} \mathcal{L}\left(f\left(\mathbf{x}^{adv}\right), y_t\right), \tag{7}$$

where y_t is the label of the class designated by the attacker.

Szegedy et al. [75] first propose an L-BFGS algorithm to transform a difficult optimization problem of finding a perceptually-minimal input perturbation into a box-constrained formulation. Many other popular methods, such as C&W [24], AdvGAN [48], and UAP [70], are also achieved by carrying out constrained optimizations. Table VI categorizes the latest constrained optimization-based attacks from the perspectives of input, attack type, and inv-metric.

Many recent works utilize constrained optimization techniques to generate adversaries. For instance, Chen et al. [143] present adGAN, a method for producing adversarial attacks that greatly degrade the performance of reinforcement learning systems. The fundamental concept of adGAN is to manipulate the current state in reinforcement learning, misleading the agent into thinking it is in a correct state, thereby causing it to make subpar decisions in each step and leading to a decrease in overall rewards. AdGAN has demonstrated its ability to transfer and adapt well to different situations. The loss function of the adversarial attack is written as:

$$\begin{aligned} \tilde{\mathbf{x}} = \max_{\mathbf{x}} \;\; & \mathcal{L}_{T_i}(f_\varphi), \\ \text{s. t.} \;\; & \|\tilde{\mathbf{x}} - \mathbf{x}^{adv}\|_2 \leq \epsilon. \end{aligned} \tag{8}$$

Here, $\mathcal{L}_{T_i}(f_\varphi)$ represents the loss function of the reinforcement learning task T_i, $\tilde{\mathbf{x}}$ symbolizes the optimal state in the Markov decision process tackled by the reinforcement learning, $\mathbf{x}^{adv}$ is the altered state, and ϵ indicates the perturbation magnitude. The aim of T_i is to learn a function f, which is parameterized by φ to maximize the expected overall discounted reward.

In the examination of the susceptibility of DL-based Radio Frequency Fingerprinting (RFFI) systems to adversarial attacks, Liu et al. [55] put forward a novel Generation Adversarial Perturbation (GAP) problem that considers the influence of actual fading channels. Furthermore, they propose a spoofing attack algorithm utilizing the S-process, outperforming the benchmark scheme in simulation tests. Moreover, Xu et al. [56] propose a graphical analysis of Radio Frequency (RF) signature perturbations after adversarial attacks. They explore the impact of fusion attacks (i.e., physical attacks coupled with adversarial attacks) on RF fingerprint classifiers from multiple perspectives.

Adversarial Transformation-enhanced Transfer Attack (ATTA) [144] is a technique that uses adversarial learning to train a CNN as an adversarial transformation network. This network is capable of capturing the most destructive deformations and transforming them into adversarial noises. The adversarial samples designed to withstand distortions induced by the adversarial transformation network are more robust and transferable. ATTA's performance may be enhanced by combining it with other transfer-based attacks, e.g., momentum iterative fast gradient sign Method (MI-FGSM) developed in [145] and Query-Efficient Black-Box Adversarial attack developed in [146]. However, overly simplistic or complex structures can negatively impact the attack's performance, as the former lacks sufficient representation power, and the latter causes the adversarial transformation network to be excessively adaptive to the backbone attack algorithm.

Luo et al. [147] propose an adversarial attack method called semantic similarity attack on high-frequency components (SSAH) based on frequency space constraints, which restricts the adversarial noise to the high-frequency components of the picture, so that the human eye perceives the noise with relatively low similarity. The framework of SSAH is displayed in Fig. 7. The attack strategy involves increasing the semantic resemblance between the adversarial sample and a randomly selected sample, while simultaneously decreasing the feature similarity between the adversarial sample and the original image. SSAH steps out of the original framework based on ℓ_p-norm constraints, and provides a new idea of adversarial noise generation and constraints in frequency

TABLE VI
CONSTRAINED OPTIMIZATION-BASED ADVERSARIAL ATTACKS

Attack	Short Description	Input	Attack type	Invisibility Metric	Strength	Weakness
GF-Attack [149]	A framework for adversarial attacks that may be launched against various GEM types.	Discrete	Black-box	ℓ_2	Good transferability on various kinds of GEMs, flexibility and extensibility, not changing the target embedding mode.	The calculation efficiency is lower than the Random method and the Degree method.
adGAN [143]	A framework that undermines the current state in reinforcement learning by enticing the agent to make sub-optimal choices in each step, leading to a reduction in overall rewards.	Discrete	White-box	ℓ_2	Model agnostic, good generalization capacity, could converge quickly in all environments, quite stable performance.	The calculation cost may be higher.
SSAH [147]	By attacking the semantic similarity of the images, a wide range of settings are applied.	Continuous	White-box	LF	More transferable across different architectures and datasets, significantly imperceptible.	No significant increase in aggressiveness.
NSGA-PSO [152]	A method for producing adversarial perturbations for digital watermarking by using an optimization algorithm.	Continuous	Black-box	ℓ_2	Satisfactory transferability across different networks, fewer queries but higher success rates.	Lower ASRs on CIFAR-10 than on ImageNet.
ATTA [144]	A CNN is trained as the adversarial transformation network through adversarial learning, allowing it to capture the most damaging deformations in response to adversarial noise.	Continuous	White-box, Black-box	ℓ_∞	The adversarial samples created are more robust and transferable. Combines well with other transfer-based attacks to boost effectiveness.	Too simplistic or complex structures reduce attack performance.
Adversarial Quantization [148]	A method specifically designed to quantify against perturbation, with the aim of minimizing quantization errors after quantization while maintaining the sample's adversarial nature.	Continuous	White-box	ℓ_2	Proposes post-processes that can be utilized for any white-box attack. Requires fewer iterations than the conventional attack process and adds little additional distortion.	Its transferability to other DNNs cannot be guaranteed. Only works on white-box attacks.
Language Model Agnostic Attack [153]	An algorithm for generating adversarial instances on a multi-model image captioning frame. The algorithm does not require language module information and controls the predicted title by attacking the visual encoder of the title frame.	Continuous	Gray-box	Meteor Score	The perturbation is calculated through a single forward pass of the deployed model, unlike the typical iterative approach, which incurs higher time consumption. The output of the language model can be controlled with no need for any information about the model.	The improvement in attack performance comes at the cost of more perceptibility of disturbances in samples.

space. However, although SSAH has a great decrease in the recognition, the attack success rate has not significantly improved. Therefore, for invisible attacks, the trade-off of their invisibility and attack success rate is still a problem that needs to be solved.

In the field of wireless signal classification, Kim et al. [3] present a proposition wherein an attacker leverages a DNN classifier to misidentify the occupied spectrum as idle. Their research illustrates that disparities in training data and channel effects between the attacker and transmitter models could considerably confine the efficacy and transferability of such attacks. Further, in a distinct study, they delve deeper into the deployment of adversarial attacks within the area of reconfigurable intelligent surfaces (RISs) in wireless systems. They demonstrate that adversarial attacks can be used in a positive manner. The introduction of adversarial perturbations to the signal could fortify covert communication by enhancing the signal detection accuracy of the intended receiver, while concurrently diminishing the detection precision of an eavesdropper. Shi et al. [50] propose a DNN-based spoofing attack to generate synthetic wireless signals that are not statistically distinguishable from the intended transmission. The opponent is modeled as a pair of transmitters and receivers, building generators and discriminators that generate the adversarial network by playing minimax games over the air. Durbha and Amuru [4] evaluate an AutoML model to classify wireless signals. Their exploration of the impacts of white-box and black-box attacks on the model provides evidence that the AutoML model's performance closely parallels that of state-of-the-art models in terms of classification, vulnerability, and transferability.

Bonnet et al. [148] propose a method specifically for quantizing adversarial perturbations. The quantization is implemented as a customizable post-processing approach that may be employed over any white-box attacks aimed at any model, with less additional distortion and fewer cycles required for the attack operation. This strategy, however, needs to access gradients available in the white-box design and does not ensure transferability to other DNNs.

Because the white-box attack needs access to predictions and labels, it is impractical for a realistic learning system. Hence, researchers have focused on black-box attacks. Chang et al. [149] present a generalized adversarial attack framework (GF-Attack). This black-box attack system can execute adversarial attacks on different kinds of graph embedding models (GEMs) without access to labels or model predictions. The objective is to improve the robustness of GEMs. Although GF-Attack has a lower computational efficiency than the Random method [150] and the Degree method [151], it can execute an adversarial attack on a range of GEM types with high transferability, flexibility, and extensibility without altering the target embedding model.

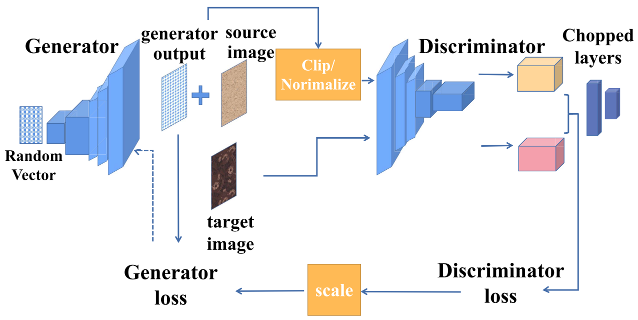


Fig. 8. Overview of the Language Model Agnostic Attack. The output of the generator combined with the source image and the target image are both inputs to the discriminator, and then the discriminator backpropagation updates the generator. The resulting generator output overlaid onto the source image to generate the title.

Non-Dominated Sorting Genetic Algorithm with Particle Swarm Optimization (NSGA-PSO) [152] is an optimization-based method for generating digital watermarking adversarial perturbations. It has higher ASRs than existing black-box attack approaches, good transferability among multiple network models, and greater resistance to image modification countermeasures. However, testing findings demonstrate that its performance is worse on the CIFAR-10 dataset, compared to the ImageNet dataset.

Despite the fact that attacks on visual models (such as CNNs) have been well studied, the adversarial vulnerability of neural image captioning has not been thoroughly investigated, because of the unique difficulties of the "multi-model" problem in subtitles. Aafaq et al. [153] suggest that images be altered in line with internal representations of visual models applied in captioning frames to deceive encoder-decoder-based image captioning frameworks. A GAN-based method is suggested, which can alter the representation of the internal layers necessary for the image in order to produce adversarial images. The diagram of the language model agnostic adversarial attack on image caption is demonstrated in Fig. 8. The attack begins by sampling a random vector from a uniform distribution via a generator. The generator output is combined with the source image. The scaled outcome is fed into the discriminator to obtain the desired depth representation. Similarly, the target image is fed into the discriminator. The discriminator back propagates the gradient to update the input image. In scale and again after deducting the original image, the disturbance is separated, and the gradient of the generator has been updated. In order to produce significant image features for the target (incorrect) class and suppression features for the source (correct) class, generators are trained to compute perturbations. The output of the generator is then attached to the source image to fabricate a title close to the title of the target image. This enables an attacker to successfully control the image's title without needing to be familiar with the caption model. But the method also has the problem that the disturbance is imperceptible and difficult to control.

Bahramali et al. [51] propose an adversarial attack against the input unknowability of a wireless communication system that is also undetectable and robust to removal. They model the potential problem as an optimization problem and solve it to obtain a perturbation generator model capable of generating a large number of input-independent adversarial sample vectors for the target wireless application. Experiments show that the proposed attack is much better than existing attacks against DNN-based wireless systems in the presence of defense mechanisms deployed by the communicating party.

All of these methods, e.g., [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153], generate adversarial examples to attack ML or DNN models by using constrained optimization, such as bi-level optimization, quantization, or particle swarm optimization (PSO). They have been shown to have good transferability spanning many network architectures, and solid resistance to example transformation defensive strategies. Most of these methods, such as GF-Attack [149] and NSGA-PSO [152], concentrate on improving the robustness of GEMs and digital watermarking adversarial perturbations and have good extension and flexibility. Moreover, they do not depend on access to the predictions and labels, making them more suitable for real-world scenarios. However, these methods could perform poorly on different datasets.

C. Gradient-Free (Heuristic) Attacks

Heuristic attacks are a sort of attack that does not depend on gradients and can include techniques such as search-based, decision-based, and drop-based methods. These methods can have their own advantages and disadvantages, and a summary of the latest gradient-free attacks from several perspectives can be found in Table VII.

Feature-Wise Convex Polytope attack (FeaCP) [154] places emphasis on limiting the placement of generated samples. It aims to find adversarial samples close to the decision boundary and correct existing areas of vulnerability in neural networks. Rather than solely focusing on the capacity for an attack, FeaCP places a greater emphasis on controlling the generation process for the purpose of defending the model. FeaCP considers the significance of adversarial instances in relation to the target model during the creation process, and provides a clear insight into the location of adversarial instances through the adversarial direction. FeaCP can be applied to other fields, such as sentiment analysis [155]. FeaCP creates potential adversarial examples with confined variables, as given by

$$\mathbf{x}^{adv} = \lambda_s \odot \mathbf{x} + \sum_{i=1}^{M} \lambda_g^i \odot \mathbf{x}_g^i \tag{9}$$

Here, $\mathbf{x}$ stands for the benign example; $\mathbf{x}_g^i$ denotes the i-th guidance example in a collection of randomly selected guidance samples of M different classes; λ_s and λ_g^i represent the tensors of coefficients that has the same shape as that of $\mathbf{x}$; $\lambda = \{\lambda_s, \lambda_g^1, \ldots, \lambda_g^M\}$ and satisfies the following condition:

$$\lambda_s^q + \sum_{i=1}^{M} \lambda_g^{iq} = 1, \tag{10}$$

where q indicates the q-th element of a tensor. Eqn. (10) ensures each feature of the composite sample is a convex combination

TABLE VII
GRADIENT-FREE ADVERSARIAL ATTACKS

Attack	Short Description	Optimizer	Input	Attack type	Invisibility Metric	Strength	Weakness
FeaCP [154]	A method to seek adversarial examples near the decision boundary.	Search-based	Continuous	White-box	ℓ_∞	Provides explainable hints on the locations of adversarial examples. Be generalizable in both computer vision and sentiment analysis fields.	Computationally expensive
AdvLB [26]	Manipulates laser beam's physical parameters for an adversarial attack.	Greedy Search based	Continuous	Black-box	ℓ_p	Direct use of laser beams as a perturbation. High flexibility to actively attack any object, even at long distances, higher temporal stability. Easy to deploy.	It is less secretive than some samples generated by other methods, such as AdvCam.
HAG [159]	Crafts adversarial examples for Hamming space search.	Search-based	Continuous	Black-box	Hamming Distance	Successfully attacks target hash models with low perceivability. High transferability at various settings, more transferable at different hash bit lengths for the same architecture.	Merging the perturbations from distinct hash digits to attack a model with a similar design did not result in a noticeable enhancement of performance.
AdvDrop [163]	Crafts adversarial examples by dropping existing information of images.	Drop-based	Continuous	White-box, Black-box	LPIPS	AdvDrop is a completely different paradigm from previous attacks and is more robust to current defense methods. Computation cost and perceptual quality are balanced.	By focusing on the frequency domain, a relatively simple strategy for eliminating information is used, which tends to lose high-frequency information.
PCAE [162]	Generates adversarial examples via principal component analysis.	DM-based	Continuous	White-box	ℓ_2	Not rely on any classifier. The impact of insufficient labeled data is limited. Competitive transferability.	PCAE has poor performance due to its high complexity in the adversarial region. Compared with neural networks, kernel PCA is limited in its ability to approximate such complex manifolds.
IoU [161]	Generates perturbations in sequence based on the predicted IoU scores from current and past frames.	Decision-based	Continuous	Black-box	Cosine Distance	The IoU score is gradually reduced by using the smallest amount of noise, which in turn reduces the accuracy of the target task. Be generalizable among different structures.	Performs less effectively than a white-box attack method, named CSA, when applied to track, as it lacks access to the trackers' network architecture.
SGADV [138]	Uses different similarity scores to generate optimized adversarial examples, effectively breaks FR-based authentication in both white-box and gray-box settings.	Similarity-based	Continuous	Gray-box, White-box	Cosine Distance, SSIM, LPIPS	High ASR and acceptable time cost.	The attack efficiency of SGADV is lower than that of FGSM, DeepFool and PGD, and no research has been carried out on its transferability.

of the bootstrap sample and the relevant features in the source samples to provide sufficient flexibility for perturbation of each feature in finding the blind spots of the DNNs.

Adversarial Laser Beam (AdvLB) is a novel attack method introduced by Duan et al. [26], which uses a greedy search and laser beams as a malicious perturbation. This method has high flexibility, allowing it to attack any object, even from long distances actively. AdvLB also has high temporal stability because of its physical attack mechanism. On the other hand, its deployment is simple, making it less secretive than other methods, such as AdvCam [156].

Adversarial attacks have also appeared in the fields of spectrum monitoring and power allocation in communications and networking [28], [47], [68], [157], [158]. For instance, Chew et al. [28] reveal that an adversarial attack, in the form of an adversarial waveform, can successfully disrupt a spectrum-monitoring system's attempts to intercept and classify signals using a CNN, demonstrating increased vulnerability as the system bandwidth grows. Zheng et al. [47] propose Primary User Adversarial Attack (PUAA) to verify the robustness of a spectrum sensing model based on DNN. PUAA incorporates carefully crafted disturbances into the benign primary user signal, significantly reducing the detection probability of the spectrum sensing model. Sun et al. [68] propose that adversarial attacks can significantly compromise the power distribution process in massive MIMO networks.

Hash adversary generation (HAG) [159] is a technique for creating adversarial examples for a search in the Hamming space that solves a widely perceived "gradient vanishing" issue[3] by introducing a smoother activation function. The objective of HAG is to generate subtly altered samples that bear no semantic connection to the original queries and whose closest neighbors come from a chosen hashing model. Even though the perceivability is still low, HAG can successfully attack target hash models. The learned perturbation is highly

[3] "Gradient vanishing" refers to the phenomenon that during the backpropagation of a deep neural network, the gradients of the network can become very small as they propagate through the layers of the network. When the gradients become too small and effectively become zero, this essentially prevents the lower layers of the network from learning any useful features [160].

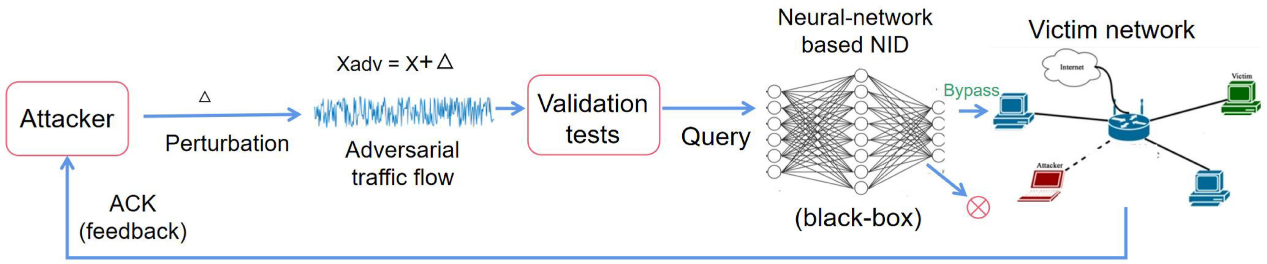


Fig. 9. An illustration of an adversarial attack against ML-based network intrusion detection models. An attacker sends several queries to the network intrusion detection model. Based on implicit/explicit feedback from the model, the attacker applies subtle perturbations to network traffic to produce adversarial traffic flow [5].

portable across settings and is more pronounced for the same architecture at varying hash bit lengths.

The importance of Network Intrusion Detection (NID) is escalating in the context of guaranteeing the availability of systems/services and safeguarding the online security and privacy of individuals. However, NIDS predicated on DNNs are not devoid of risks. A recent study [5] presents TIKI-TAKA, a novel framework designed for adversarial attacks against NIDS built on DNN. The authors trained three state-of-the-art DNN models, i.e., Multilayer Perceptron (MLP), CNN, and Conv-long short-term memory (C-LSTM), on publicly accessible datasets, employing five classes of adversarial decision-based attacks to disrupt the models. As shown in Fig. 9, an attacker might send a traffic stream to the target network, which would first be checked by the network intrusion detection models. The attacker will then adjust and apply subtle perturbations to the malicious traffic based on the feedback, resulting in adversarial samples that can eventually compromise the effectiveness of the NIDS. Experimental results underscore that while DL-based NIDS exhibit a high detection rate, they remain susceptible to adversarial samples.

An Intersection over Union (IoU) attack [161] is a black-box, decision-based technique for visual object tracking. It creates disturbances using the calculated IoU scores from current and prior video frames. The attack decreases the accuracy of temporally consistent bounding boxes by lowering the IoU scores. Denote the original example (i.e., the original image in the video frame) as $\mathbf{x}$, the heavy noise example as $\mathbf{X}$, and the intermediate example on the i-th iteration as $\mathbf{x}^{(i)}$. The IoU attack labels η as a nearby assumption based on $\mathbf{x}^{(i)}$ and advances $\mathbf{x}^{(i)} + \eta$ towards the highly noisy example $\mathbf{X}$ by following the update rule below:

$$\mathbf{x}^{(i+1)} = \left(\mathbf{x}^{(i)} + \eta\right) + \alpha \cdot \psi\left(\mathbf{X}, \mathbf{x}^{(i)} + \eta\right), \tag{11}$$

where α represents the stride towards $\mathbf{X}$ and $\alpha \cdot \psi(\mathbf{X}, \mathbf{x}^{(i)} + \eta)$ represents the disturbance in the direction of greater noise, i.e., in the direction of the normal to the noise level contour.

AMGmal [80] constitutes a new technique for generating adversarial examples. This method, aiming to deceive malware detectors predicated on DNNs while minimizing the requisite amount of perturbation, maintains its efficacy even when defensive mechanisms are operational.

To circumvent the constraints of model-dependent approaches, such as the C&W constraint undergone by a well-trained classifier in [24], Zhang et al. [162] present the Principal Component Adversarial Example (PCAE) method. PCAE produces adversarial samples without a specific target in mind. It is based on the idea of the adversarial zone where data points offer a possible danger to all classifiers. As an untargeted adversarial sample generation approach, PCAE utilizes a data manifold that does not depend on classification models. As a consequence, it is immune to overfitting and the restrictions of inadequate labeled data.

Ye et al. [46] assess the performance of various white- and black-box adversarial attack algorithms on OFDM detectors. This work reveals the remarkable efficiency of adversarial attacks in impairing system performance, underscoring the merits of the Virtual Adversarial Method (VAM) and Zero-Order Optimization (ZOO) attacks in white-box and black-box contexts, respectively.

Different from all previous attacks, AdvDrop is a novel adversarial attack proposed by Duan et al. [163], which creates adversarial examples by removing certain features from benign images. This makes the resultant images unnoticeable to humans but essential for DNNs to misclassify them. AdvDrop is more resistant to existing defensive mechanisms, e.g., AT [164] and feature squeezing [165], and paves the way for a new approach to assessing the robustness of DNNs. Focusing on the frequency domain, it deletes high-frequency information more often than low-frequency information.

Moreover, Sun et al. [68] present a new adversarial attack framework, namely generating practical malicious traffic (GPMT), which is designed to generate adversarial traffic capable of deceiving ML-based traffic detection systems, as shown in Fig. 10. This framework offers heightened efficiency and generalizability, manifesting substantial evasion growth rates across diverse models and datasets.

In response to the issue of malicious traffic, Yang et al. [79] introduce a novel traffic obfuscation methodology, namely traffic obfuscation adversarial network, namely TONet, predicated on the employment of adversarial neural networks to generate disturbance vectors. The obfuscation samples generated via this approach exhibit an exceptionally low disturbance cost and an exceedingly high defense success rate (i.e., $\geq 99\%$) in scenarios involving known adversaries. This methodology demonstrates robustness against unknown models for adversary attacks and optimizes both computational complexity and implementation speed.

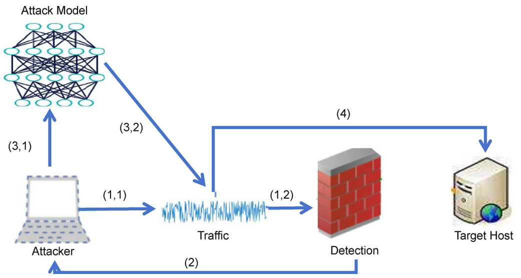


Fig. 10. The adversarial attack process for ML-based malicious traffic detection: (1.1) The attacker attempts to attack the target; (1.2) The attacker generates original malicious traffic; (2) The detection model detects malicious behavior and issues an alarm to block network traffic; (3.1) The attacker receives feedback and builds the adversarial attack model; (3.2) The attack model modifies malicious traffic until it can successfully escape detection; and (4) adversarial traffic escape detection successfully [68].

The majority of adversarial attack strategies rely on label data, but face recognition (FR) authentication systems don't keep track of the label data for the target user. A similarity-based gray-box adversarial attack (SGADV) is put forth by Wang et al. [138] to address the shortcomings of current adversarial attacks on FR authentication systems. To implement adversarial attacks based on benchmark labels, a conditional binary cross-entropy (C-BCE) objective function is also designed as a baseline against FR-based authentication. Additionally, the experimental findings demonstrate that the pre-trained model is not secure in practice even if the database for face template storage is unharmed, demonstrating the importance of this research for raising the privacy threat to users. SGADV achieves effective attacks and a satisfactory time cost, but it is less efficient than FGSM [54], PGD [22] and DeepFool [76], and no studies have been conducted on transferability.

The above-mentioned attack methods, i.e., FeaCP [154], AdvLB [26], HAG [159], PCAE [162], IoU [161], AdvDrop [163] and SGADV [138] represent the SOTA gradient-free attacks for DNNs.

- FeaCP [154] defends neural networks by limiting generated samples, finding close adversarial samples, and controlling the generation process while providing insight into their location.
- AdvLB [26] is a new attack method that uses a search method and laser beams to attack objects from a distance. It is easy to use but not as secretive as other methods.
- HAG [159] generates adversarial examples in the Hamming space with no semantic connection to the original queries. Its perturbation is portable across settings.
- The IoU attack [161] is a method that uses past and current frames to reduce the accuracy of object tracking by generating perturbations.
- PCAE [162] is an approach for crafting adversarial examples that do not require a target and do not have the issues of overfitting or lack of data.
- AdvDrop [163] generates adversarial examples by modifying image frequencies, making it challenging to defend against and a useful tool for testing DNNs' robustness.
- SGADV [138] utilizes different similarity scores to generate optimized adversarial samples, effectively breaking FR-based authentication in both white-box and gray-box attacks.

All of these methods have their own strengths and limitations. For example, AdvLB is highly temporally stable but less secretive. PCAE is target-free, but it does not rely on any classifier, so it is hard to evaluate its performance. AdvDrop is more robust to current defense methods, but it is relatively simple and focuses on the frequency domain.

D. Adversarial Patch

Although adversarial sample attacks, such as PGD [22] and Contrastive AT [139], can achieve a high ASR and undetectable perturbation effect, its generalization ability is generally poor to be used in the physical world, and a specific perturbation must be generated for each attack. As a result, adversarial patch attacks [26], [166] come into view as a variant of the adversarial sample attack, in contrast to adversarial sample attacks where an attacker always aims to minimize the level of perturbation to avoid detection. In adversarial patch attacks, the attacker never again confines themselves to imperceptible changes. The attack generates an image-independent patch, which can be set anyplace in the image to attack a DNN-based image classifier and cause it to output a specified target class. Table VIII collates the latest adversarial patch methods.

The advantage of adversarial patching over adversarial sample attacks is that adversarial patching can be more targeted and effective in deceiving a deep learning model by creating a small, localized patch that can be placed at a specific location of an image. By contrast, adversarial sample attacks typically add noise or distortion to an entire image. Moreover, adversarial patching can potentially attack deep learning models in real-world scenarios where digital attacks are impossible, such as in physical security systems. This makes it a powerful tool for attackers to bypass ML-based security systems in many practical application scenarios. However, adversarial patching requires more effort and knowledge to create and raises important questions regarding the responsible development and deployment of ML systems.

FaceAdv [167] is a physical adversarial attack technique, which uses malicious stickers to fool face recognition applications. FaceAdv comprises a malicious sticker generator and a converter. The generator makes a variety of differently-shaped stickers (some examples of stickers are shown in Fig. 11). At the same time, the latter applies the stickers digitally on human faces and provides the generator with attack results to enhance its efficacy. Despite changes in environmental conditions, FaceAdv can dramatically increase the success rate of avoidance and simulated attacks, showing robustness. However, the "sticker" attacks require a high degree of hardness to avoid early detection by humans. They also require the adversaries to physically access the target object for pasting the stickers, which may not always be possible.

Robust Physical Perturbation (RP2) [169] is a generic attack algorithm that produces perturbations robust to varying angles and distances under different physical conditions. The perturbations are visible but inconspicuous and only perturb objects (e.g., road signs) without disturbing the object's environment.

TABLE VIII
ADVERSARIAL PATCHING METHODS

Strategy	Brief description	Performance
FaceAdv [167]	A physical attack that creates adversarial stickers to trick face recognition systems, made up of a sticker generator and a converter.	Keep robustness in dodging and impersonating attacks. It does not affect the performance of the face detector. Good transferability in both the digital and physical worlds.
RP2 [169]	Samples from a distribution that simulates physical dynamics to project calculated perturbations into a graffiti-like shape.	Perturbations that are robust against widely varying distances and angles can be generated under different physical conditions.
PS-GAN [170]	Perceptual sensitivity is used to improve the visual rationality and aggression of adversarial patches. A visual attention mechanism is employed to capture the sensitivity of spatial distribution.	Guides the attack positioning of the adversarial patch for stable attack effects. Taking it a step further, PS-GAN can generate adversarial patches instantly.
MultiD-WGAN [171]	Based on the idea of generating adversarial patches by GANs, the data-driven MultiD-WGAN is proposed, which can simultaneously enhances the offensive power and authenticity of adversarial patches by multiple discriminators.	Simultaneous enhancement of the aggressiveness and authenticity of adversarial patches through multiple discriminators.
Bias-based Framework [172]	Takes advantage of perception bias and attention bias to improve attack capabilities.	The resulting adversarial examples allow for greater transferability between different models.
Singular image based [174]	Determines adversarial patch's location according to the perceived sensitivity of the victim model, encouraging patch alignment with background images through AT.	It has strong attack ability in a white-box environment and good transferability for black-box environments, which is more difficult to detect and can also be applied to the physical world.
AdvCam [156]	The style shift approach is used to achieve concealment, and adversarial strength is achieved using the technique of adversarial attack.	AdvCam's forged adversarial samples in the digital and physical worlds are highly stealthy and still valid when it comes to spoofing the latest DNN-based image classifiers
MAP [177]	MAP is optimized using CSM and ME losses.	Successfully attack multispectral personnel detectors in both physical and digital spaces.
Multi-perspective Environments [178]	Considers the effects of viewing angle changes in multi-perspective environments by integrating adversarial patches with perspective geometry transformations.	Viewing angles have a strong impact on the effectiveness of adversarial patches. In some scenarios, adversarial patches lose most of their effectiveness, opening up new opportunities for adversarial defense.

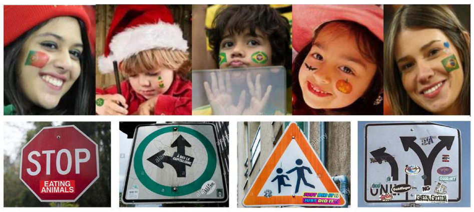


Fig. 11. Examples of stickers for faces and traffic signs [168].

The algorithm utilizes a mask to transform the estimated perturbations into a graffiti-like form after sampling from a range of simulated physical dynamics. An attacker may then print out the resultant perturbations and apply them to the road sign under attack, resulting in a high rate of misclassification of the target by the road sign classifier, which might lead to catastrophic consequences.

Perceptual sensitivity is a crucial aspect of visual recognition systems. The more natural-looking the generated adversarial blocks, the more likely the attacks are successful.

Perceptual-Sensitive Generative Adversarial Network (PS-GAN) [170] uses perceptual sensitivity to improve the visual plausibility and attack capability of adversarial patches. It adopts a visual attention mechanism to capture the sensitivity of the spatial distribution and guide the localization of the adversarial patches for a stable attack effect. PS-GAN can also generate adversarial patches on-the-fly without the need to access the target model at the time of inference. This makes it a powerful tool for attackers looking to bypass visual recognition systems in real-world scenarios. Similarly,

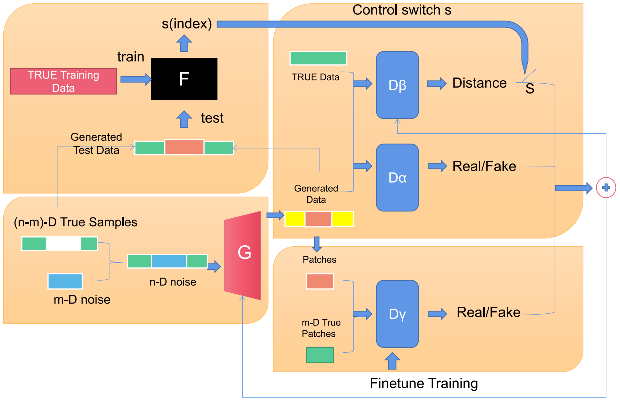


Fig. 12. The framework of the MultiD-WGAN method. It is composed of a generator G, several discriminators, and a target classifier F.

Wang et al. [171] propose a data-driven, Muti-Discriminator Wasserstein GAN (MultiD-WGAN) algorithm based on GANs to craft adversarial patches that focus on the perceived sensitivity of the attacked neural network model, as shown in Fig. 12. The algorithm enhances both the aggressiveness and authenticity of adversarial patches by utilizing multiple discriminators. The research demonstrates, theoretically and experimentally, a positive correlation between attack strength and attack capability.

Wang et al. [172] devise a bias-based framework to produce generic adversarial patches that exploit attentional bias and perceptual bias to improve attack capabilities and increase the generality of adversarial patches. The framework uses style similarity [173] to extract a patch that comes before texture from a hard example with high model uncertainty and accounts for perceptual bias. An attentional bias is utilized by obscuring the same attentional patterns shared by models, which are identical for the same image across multiple models. This allows the created adversarial patch to be more transferable between models.

To tackle the hindrance of feeble disguise of adversarial patches and lengthy computational time, Bai et al. [174] advance a procedure to bring forth inconspicuous adversarial patches exploiting singular images. The technique initially ascertains patch areas depending on the perceptual sensitivity of the target model, and then fabricates adversarial patches in a coarse-to-fine system, which utilizes multiscale producers and judges. The patches are urged to coordinate with the background image through adversarial training. At the same time, it still maintain a powerful attack aptitude. Experimentally, the proposed method has demonstrated formidable attack capability in white-box settings and good transferability for black-box situations, making it difficult to identify.

To increase adversarial stealthiness and camouflage flexibility while maintaining adversarial strength, AdvCam [156] uses a style migration approach to achieve stealthiness and an adversarial attack technique to strengthen adversarial capabilities. The attacker specifies the target image, the target attack area, and the intended target style. AdvCam transforms significant adversarial perturbations into adjusted styles. The latter is then disguised in the target object or the background outside the target. Experiments conducted in both the digital- and physical-world scenarios show that AdvCam's faked adversarial samples are highly concealable and yet still effective in spoofing the latest DNN-based image classifiers.

The existing studies on adversarial patches, such as those developed in [175], [176], have only looked at the robustness of single-spectrum (RGB or Thermal) models and have not evaluated multispectral models. They have only analyzed digital space perturbations and have not considered vulnerabilities in the physical world. To address these limitations, Kim et al. [177] introduce a new framework for generating multispectral adversarial patches (MAP) using material emissivity (ME) loss optimization and cross-spectral mapping (CSM). The experiments show that the generated MAP can successfully attack multispectral personnel detectors in both physical and digital spaces, highlighting the need for further research in this area. Tarchoun et al. [178] investigate the influence of viewpoint on the efficacy of adversarial patches. In order to replicate the effect of perspective alterations in multi-perspective settings, they combine known adversarial patches with perspective geometric transformations. Experiments demonstrate that perspective substantially affects the efficacy of adversarial patches, which can sometimes drop substantially. This finding encourages academics to investigate the influence of viewpoint on adversarial attacks and reveals new options for adversarial defenses.

The above-mentioned methods represent the current SOTA in adversarial patch generation, with a focus on image recognition systems. These methods include FaceAdv, which crafts adversarial stickers to deceive facial recognition systems [167]; RP2, which produces perturbations that are robust to varying distances and angles under different physical conditions [169]; PS-GAN, which uses perceptual sensitivity to improve the visual plausibility and attack capability of adversarial patches [170]; MultiD-WGAN, which enhances both the aggressiveness and authenticity of adversarial patches by utilizing multiple discriminators [171]; AdvCam, which increases adversarial stealthiness and camouflage flexibility while maintaining adversarial strength by using a style migration approach and an adversarial attack technique [156]; and MAP, which generates multi-spectral adversarial patches to attack multi-spectral personnel detectors in both physical and digital spaces [177].

Some of the methods, i.e., FaceAdv [167], RP2 [169], and PS-GAN [170], have demonstrated considerable strengths. For example, they are powerful tools for attackers looking to bypass ML-based security systems in the real world. They improve the visual plausibility and attack capability of adversarial patches. They can increase adversarial stealthiness and camouflage flexibility, while maintaining adversarial strength. They consider the impact of perspective in adversarial attacks.

On the other hand, they also have some weaknesses. Some methods require a high degree of hardness to avoid early detection by humans. Some methods require the attacker to physically access the target object to paste stickers, which may not always be possible. Moreover, they have not been thoroughly tested in multispectral models. There is still much work to be done in terms of understanding the limitations and weaknesses of these methods, and developing effective countermeasures to protect against them.

E. Transferability of Adversarial Attacks

Transferability accounts for the ability of adversarial attacks to be applied to different models and datasets. Ideally, from the perspective of attackers, an adversarial sample generated to deceive a specific model can also deceive other models. This is significant because it means that an attacker does not need to generate a new adversarial sample for each model or dataset it wants to attack, which can be time-consuming and computationally expensive. Suppose an adversarial sample is highly transferable. In this case, it is more likely to be successful in the real world, where the attackers might not know the specific model or dataset they are trying to attack. This makes the attack more powerful and can increase the ASR.

Goodfellow et al. [54] believe that a linear model is sufficient to produce adversarial instances in high-dimensional space, rather than relying on highly nonlinear features of DNNs. They explain that the reason for cross-model generalization is that adversarial examples are highly consistent with the weight vector of the model, and different models that carry out the same task learn similar functions. Su et al. [179] evaluate eighteen DNN-based image classification models and concluded that untargeted attacks obtain higher transferability than targeted attacks. The transferability of adversarial samples was sometimes symmetric. They also discover that most adversarial samples from one model could only migrate among similar models. In addition, the transferability of the Visual

Geometry Group (VGG) model [180] performs far better than that of other models, making it a solid starting point for enhancing the black-box transfer-based attack.

According to [162], the transferability of adversarial cases is mostly due to the junction of adversarial area divisions and distinct classifier borders. The authors propose PCAE, a data-generated approach that can generate more transferrable adversarial instances than certain model-dependent methods. They also demonstrate that the target-free strategy might discover more transferrable adversarial scenarios and that target-free adversarial instances have greater transferability when model and/or dataset similarity is high.

Many other researchers have made efforts from various directions to strengthen the transferability of adversarial cases. AoA [129] focuses on attention heat maps, for which diverse DNNs provide comparable results, making AoA highly transferable. Contrastive AT [139] provides adversarial instances on ensemble models as opposed to a single model and has shown its efficacy in the black-box situation for improving transferability.

Wang et al. [138] study the impact of proxy representation learning on the transferability of adversarial attacks in gray-box graphs. The authors put forth that the proxy models need to maintain the consistency of node topology in the embedding layer and input layer, and use the SRLIM to maintain the topology of nodes mapped from a non-input space to a Euclidean embedding space. The proposed method realizes the improvement of the generalization and transferability of adversarial attacks.

FaceAdv [167] use the collection of face recognition systems to train the sticker generator and update the loss function. ATTA [144] enhances the transferability of generated adversarial samples by adversarial transformations, which is a network of adversarial transformations that automates the distortion adjustment procedure. IGA [133] is a transferrable attack for unknown link prediction approaches. In IGA, since the perturbations caused by GAE are universal and the attack is transferrable, the adversarial graph may still be successful in a variety of link prediction models. This is due to the fact that GAE can extract critical information from graphs in pursuit of link prediction.

As discussed above, several studies have been undertaken on the transferability of adversarial instances, or the ability of an adversarial attack to deceive different models or datasets. Studies have shown that the transferability of adversarial examples depends primarily on the closeness of the models or datasets under attack, and that untargeted attacks tend to have more transferability than targeted ones. Some studies have suggested approaches to enhancing the transferability of adversarial instances, including the use of ensemble models, attention heat maps, and adversarial transformations.

However, there is still room for improvement in the transferability of adversarial examples, particularly in creating more effective and efficient transferable attacks and in better understanding the underlying causes of transferability. Moreover, current methodologies tend to focus on image classification models. There is a need for more studies on other types of models, such as NLP. For example, Wallace et al. [181] build an imitation model like the victim model to study the transferability of a black-box machine translation system by using gradient-based attacks.

On the other hand, it is important to develop new approaches to defending against transferable adversarial attacks. One strategy is to use transferable adversarial examples to enhance the robustness of DNN models through adversarial learning, as suggested by [182]. The limitation of this strategy is that it requires a large number of transferable adversarial examples for training, which can be time-consuming and can adversely affect the prediction accuracy of the DNN model on natural examples due to an increased ratio of adversarial examples in the training dataset. Another strategy is to assemble various defenses into an ensemble solution to compensate for the lack of diversity in a single defense mechanism. For example, Deep Fusion Defense [183] employs three or five DNN models trained with different perturbation magnitudes to achieve superior performance in defending against transferable adversarial examples. However, the ensemble strategy can worsen the time and computational cost of the defense. Therefore, developing few-shot (i.e., using fewer training examples) solutions for defending against transferable adversarial examples is essential.

Recently, Zhou et al. [184] indicate that introducing randomness into neural network models can hinder the transferability of adversarial attacks. They also reveal that the transferability of adversarial attacks is closely related to the spread of DNN models distributed in the version space and the severity of adversarial attacks. As a result, the robustness of a DNN model can be enhanced using any subset of the DNN models, or by adding a mild Gaussian noise to the weight of the pre-trained model. In addition, the robustness of adversarial ensemble training also has great potential for improvement combined with randomization techniques.

Nowroozi et al. [72] propose two current defense mechanisms to prevent the transferability of adversarial attacks. The first approach is to fine-tune the classifier with the most powerful attacks (MPAs) each time a shift occurs against a given adversarial attack. Another strategy relies on using the long short-term memory (LSTM) [185] architecture, instead of CNN, as the target network, allowing the attacker to have less attack information than a previous system that did not know the target network architecture. Moreover, the Luring Effect [186] is a new way to boost the robustness of DNN models against black-box transfer attacks. The key concept is situated in conventional network security methods based on deception, which does not need a labeled dataset but needs access to the target model's predictions. Some other defense methods, such as Robust Soft Label Adversarial Distillation (RSLAD) [187] and Dual-Domain based Defense (D2Defend) [188], demonstrate their effectiveness in defending against transfer-based black-box attacks.

F. Summary and Lessons Learned

Adversarial attacks can be launched in several different ways, including gradient-based, optimization-based, and search-based methods. Gradient-based attack schemes are

TABLE IX
PERFORMANCE OF ADVERSARIAL ATTACK METHODS IN COMMUNICATIONS AND NETWORKS

Attack	Effectiveness	Imperceptibility	Complexity	Transferability	Impact on Communications and Networks
PGM [51]	✓✓✓	✓✓✓	✓✓✓	✓✓	DNN-based wireless communication systems are vulnerable to adversarial attacks even when employing well-considered defenses [201], and call into question the employment of DNNs for a number of tasks in robust wireless communication.
AMGmal [80]	✓✓✓	✓✓✓	✓	✓✓	AMGmal finds an optimal balance between maximum escape rate and minimum perturbation amplitude. It can be generalized to other attack methods as a general post-processing method to minimize perturbation [202], [203].
Heatmap attack [52]	✓✓	✓✓	✓✓✓	✓	Payload-based IoT identification solutions [204] have flaws that attackers can exploit to evade IoT identification solutions [24]. The proposed attack will help evaluate the robustness of defense models.
Adversarial waveform [28]	✓	✓✓	✓✓✓	✓	Adversarial attacks, in the form of adversarial waveforms, can successfully disrupt spectrum monitoring systems' attempts to intercept and classify signals using CNNs, showing a vulnerability that increases with bandwidth [205], [206].
TONet [79]	✓✓	✓	✓	✓✓✓	In order to solve the privacy leakage problem in communication with communication pattern-based analysis [207], [24], TONet provides an efficient traffic obfuscation method based on neural networks, which generates traffic distortion with minimal overhead and computational cost.
GAP [55]	✓✓✓	✓✓	✓	✓	To solve the problem that existing adversarial attack schemes ignore the influence of the actual fading channel between the attacker and the sensor [208], [209], GAP provides a new adversarial attack scheme based on convex programming.
GPMT [68]	✓✓✓	✓	✓	✓✓	The adversarial attack framework GPMT can generate actual adversarial traffic to mislead ML-based detection [210], [211]. With little prior knowledge but more adversarial and practical examples, it is more effective and versatile than other methods.
Channel-aware adversarial attacks [44]	✓	✓	✓✓	✓	By taking into account the channel effect of the opponent on each receiver, the classifier of the counter disturbance is made to deceive different receivers [205], [206]. An authentication defense method based on random smoothing is introduced, which uses noise to enhance training data to make the modulation classifier robust.
Surrogate model based attack [3]	✓	✓	✓	✓	The performance of adversarial attacks against a wireless signal classifier heavily relies on the reliability of a surrogate model that depends on the difference of channels experienced by the adversary and the transmitter [212], [213].

known for their high ASRs and good transferability. They still have limitations, such as high computational and time costs, as well as the issue of "gradient saturation", which reduces their effectiveness [189]. Moreover, gradient-based methods are relatively easy to defend against [190]. Many existing defenses, such as obfuscated gradients [191], can effectively block most gradient-based attacks.

Constrained optimization-based attack methods have good transferability but are also known for their high computational and time costs, making them difficult to use in time-sensitive applications [164]. Search-based attack methods are highly transferable and can be extended to domains beyond image classification [12]. However, for more complex data sets, searching for the optimal adversarial sample needs more iterations and high computational costs, and it is difficult to find the appropriate search start point. Currently, search-based methods are mainly applied to the optimization of other adversarial sample generation algorithms [192].

Adversarial attacks can be performed not only on images but also on other types of media, such as audio [193], [194], [195], text [196], [197], and wireless signals [44], [46], [60]. CNNs normalize all inputs to continuous signals, regardless of their semantic meanings. The major difference between images and other types of media is in their dimensions, which require adapted convolutional kernels for feature extraction. In this sense, the same adversarial attacks or their variations are largely applicable to inputs with media other than images.

Adversarial examples affect classification tasks and threaten other deep-learning tasks, such as regression. For instance, a neural network that solves the power allocation problem for a massive MIMO system can be misled by FGSM [54], PGD [22], or UAP [70] attacks, which were originally developed for image classification problems [198], [199]. Adversarial attacks, such as FGSM, I-FGSM [74], and PGD, have also shown effectiveness in linear regression tasks [200]. To this end, research progress made on adversarial attacks and defenses, e.g., regarding the image classification tasks, can be highly beneficial to other types of deep learning tasks.

Future investigation is expected to focus on reducing attack costs, improving transferability across different datasets and models, and extending to more deep learning domains. Additionally, it is important to strike a balance between perturbation visibility and attack success in order to develop effective adversarial attack methods.

As presented in Table IX. The methods have now been ranked based on four parameters: invisibility, efficiency, portability, and computational complexity, with check marks meaning better performance.

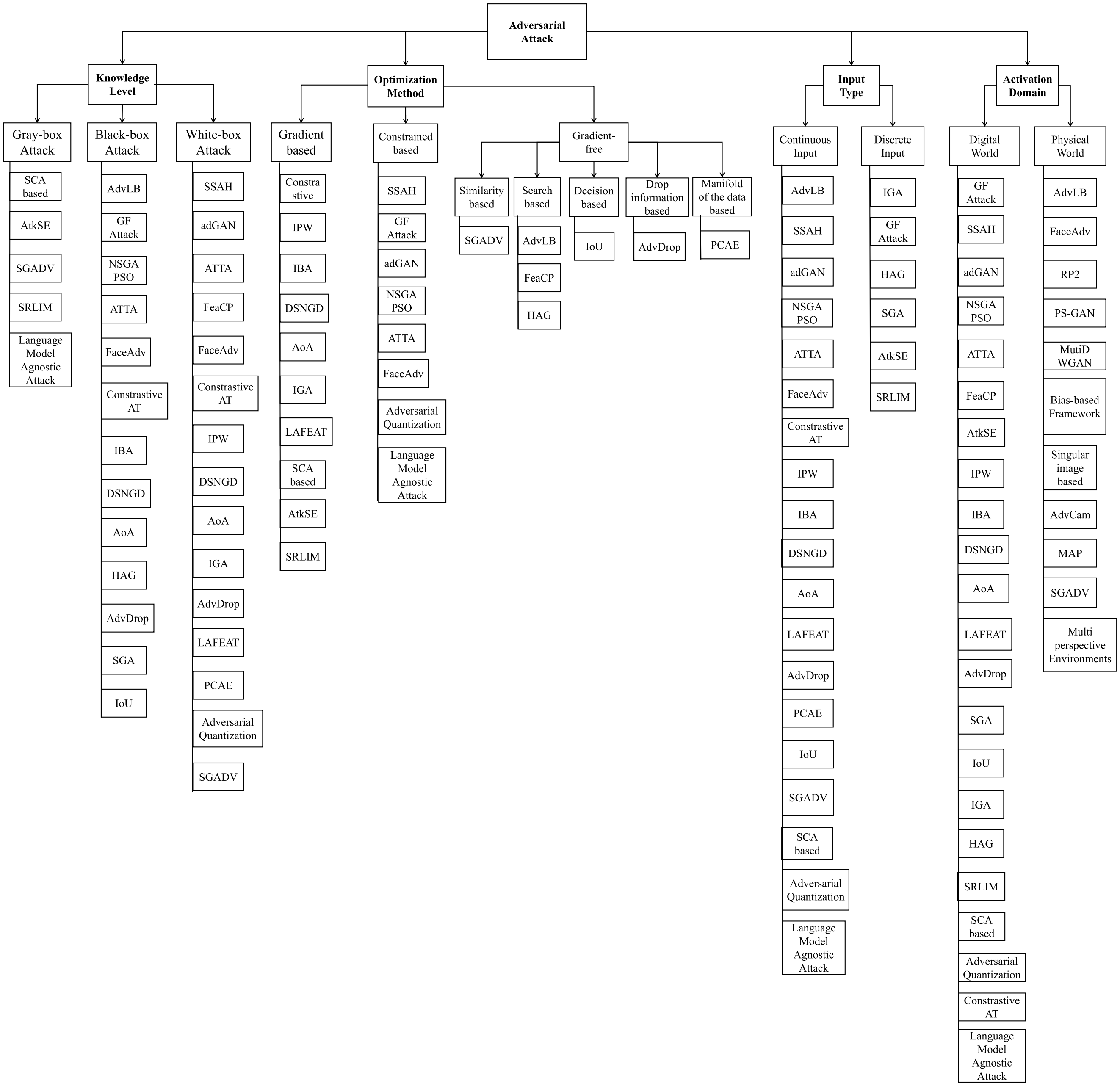


Fig. 13. Anatomy of breakthroughs in adversarial attacks since 2021.

IV. State-of-the-Art Adversarial Defense Techniques

To counteract adversarial attacks, various adversarial defense techniques have been devised. These techniques are designed to counteract specific attack techniques and range from specific defenses to general defense strategies. Deep learning models need to have the ability to counteract such attacks to maintain their accuracy and effectiveness.

A. Overview

Typical adversarial defense techniques that have been developed include:

- *Adversarial Learning:* Adversarial learning is a type of deep learning technique that involves training a model to improve its robustness against adversarial examples. One of its key techniques is adversarial training, which involves adding adversarial samples to the training process to improve the robustness of a DNN model. By continuously learning the features of adversarial samples, the model can better defend against attacks that involve adding subtle disturbances to input samples. This can improve the accuracy and effectiveness of the model in many real-world scenarios, where it may be exposed to adversarial samples designed to "trick" it into making incorrect predictions.
- *Monitoring:* This is a strategy for identifying adversarial samples, which are input samples modified to cause a deep-learning model to make incorrect predictions. To monitor for adversarial samples, special models can be

set up at key points in the system to identify these samples and provide early warning of potential adversarial attacks. This allows the system to take proactive measures to defend against attacks and maintain the integrity of the model's predictions.

- *Model Robustness Design:* This involves using specific filtering structures in the model to enhance its resilience against adversarial noise. Adversarial noise refers to the subtle disturbances that are attached to input samples to incite the model to produce false predictions. By designing the model to be more resistant to adversarial noise, it can better defend against these types of attacks and maintain the accuracy of its predictions.
- *Adversarial Perturbation Structure Destruction:* This involves using various strategies to attenuate the effect of adversarial noise and prevent attacks on the deep learning models. The strategies may include the use of filtering algorithms, noise structure destruction algorithms, and noise coverage algorithms in data stream processing. The goal is to achieve more resilience to adversarial noise, which is the subtle disturbance added to input samples to cause the model to make incorrect predictions.

These four aspects are important to build robust, secure, and resilient deep learning systems, particularly in fields where the integrity and accuracy of the model's predictions are critical, e.g., in network security and finance.

With the constant emergence of new and increasingly destructive adversarial attack methods, many research efforts have been devoted to exploring corresponding defenses. Current adversarial defense strategies can be categorized into two prevalent strategies: One strategy is based on detection and data preprocessing, and the other strategy improves adversarial robustness.

From a DNN model perspective, adversarial learning can be interpreted as gradient masking, which can refer to a class of techniques that hide model gradients from adversaries and prevent the adversaries from obtaining the correct gradients of the models, such as Graph Adversarial Training (GraphAT) [214] and Robust CNN Training [215], as will be delineated in Section IV-C4. More generally, gradient masking can refer to the outcome or effect of techniques designed to take other approaches (e.g., defensive distillation [187]) to defend against adversarial attacks and resulting in obscured gradients of the network models under attack.

B. Adversarial Attack Detection and Data Preprocessing

This type of defense method primarily detects adversarial attacks through technical means and pre-detected adversarial samples, or preprocesses the input data and destroys some key structures that constitute the adversarial samples.

1) Adversarial Attack Detection: As an adversarial defense method, adversarial sample detection has also attracted much attention from researchers. Given a sample, the goal is to directly detect whether it presents a threat. In essence, the detector is trained on both the raw and adversarial sample datasets to identify adversarial samples by measuring the differences between them caused by the adversarial perturbation [216]. A prominent detection method, H&G [217], utilizes techniques, e.g., PCA, softmax, and the reconstruction of adversarial images. These methods exploit the differences between original and perturbed images, but can be easily bypassed by attacks that target at them.

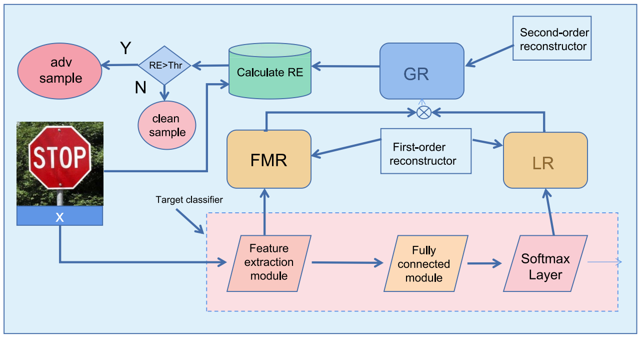


Fig. 14. The CMAG workflow during the deployment phase. The CMAG is made up of FMR, LR, and GR, which are three reconstructors. When the reconstruction error exceeds the stated threshold, the sample is considered hostile.

An algorithm called RObust SAliency (ROSA), presented by Li et al. [218], is an innovative technique for enhancing the robustness of FCN-based salient object recognition models against adversarial attacks. It works by adding universal noise to the input image, then using a two-part system to predict the saliency map of the image: A piecewise-masked component that disrupts adversarial noise patterns while preserving boundaries, and a context-aware refinement component that adjusts the saliency mapping by using contrast modeling. ROSA enhances network robustness to attacks and performs comparably or better on natural images than current methods, which only focus on non-target attacks. The defensive performance against target attacks has yet to be explored.

Aiming to enhance NIDS, Debicha et al. [83] develop an effective adversarial detector predicated on transfer learning of DNNs. They propose that in scenarios involving parallel intrusion detection system (IDS) designs, harnessing the synergy of multiple detectors can markedly augment the detectability of adversarial traffic, outperforming a singular detector. Correspondingly, within the domain of IoT intrusion detection, Jiang et al. [84] introduce a novel framework titled Feature Grouping and Multi-model Fusion Detector (FGMD). This framework fortifies defenses against adversarial attacks through the strategic grouping of features and fusion of multiple models.

Cascade model-aware generative (CMAG) [219] is an adversarial sample detection technique that consists of two first-order reconstructors, including a Feature-Map Reconstructor (FMR) and Logit Reconstructor (LR), and a second-order Global Reconstructor (GR). Rebuilding the logit and feature mapping produces an interpretable representation of the final convolution layer. If the reconstruction error (RE) of a sample relative to GR exceeds the predetermined threshold, the sample is classified as adversarial. The process of a CMAG during deployment is shown in Fig. 14. CMAG provides a new means to detect the presence of adversarial samples, which can accurately detect high-quality adversarial samples compared to existing generative model-based detection methods, e.g., Fence FGAN [220] and UADD-GAN [221]. The drawback of CMAG is that it utilizes a simplistic autoencoder as the generative model, and may not yield satisfactory results for complex datasets.

Zhang and Wang [222] state that adversarial attacks primarily attain their objectives by altering pixel values and that such attacks often insert perturbations in regions with high textures. In response, they presented a step-based deep learning network known as ADNet. ADNet is a DNN model for adversarial example detection by using steganalysis and attention mechanisms. It features an attention module, an adversarial attack attention module (AAAM), which pays additional attention to vulnerable parts throughout the process of feature learning, hence increasing the model's accuracy. To reduce the misclassification of regular samples in the detection phase, a special adversarial loss function has been designed to fine-tune the model, resulting in impressive outcomes. As an end-to-end model, ADNet does not rely on the extraction of high-quality features, hence reducing the cost of human participation. However, it encounters the problem that the detection rate for adversarial samples is better than the classification accuracy for clean samples.

Addressing the potential security risks adversarial attacks posed to ML-based IDS, Li et al. [6] conduct a detailed examination of adversarial attackers' ability to deceive detectors used in IIoT, specifically EIFDAA. The robustness of IDS is significantly improved through adversarial training. The improved IDS effectively resist adversarial attackers while preserving the original detection rate of attack samples.

Freitas et al. [223] indicate that adversarial vulnerability is a consequence of the excessive sensitivity of a model to good generalization features in the data. Since the model not only learns robust features but also information about non-robust features during training, models can be vulnerable despite maximizing accuracy. In light of this, the concept of adversarial vulnerability is extended to combine with prior human knowledge, and a new approach named UnMask is proposed, which is a framework for detection and protection against adversaries that relies on strong feature alignment. UnMask quantitatively evaluates the resemblance between the extracted and expected features, selects an adversarial perturbation to detect with a given similarity threshold, and protects the model by predicting the correct class that best fits the extracted features. The method highlights the advantage that even if an attacker can manipulate the predicted class labels by slightly changing the pixel values, simultaneously manipulating all the individual features that make up the image together is a more challenging task. Currently, UnMask only focuses on non-target attacks, and the defensive performance against target attacks has not been validated.

Although most adversarial defense detection methods have offered satisfactory results, some problems are yet to be addressed, including excessive reliance on target models, difficulty in resisting transfer attacks, relatively weak generalization capabilities, and so on. Model-independent methods to detect adversarial inputs are developed by Wang et al. [224]. The architecture is reviewed in Fig. 15. As the primary architecture of the detector, an LSTM network is trained to capture variations in the logit sequence distribution. They examine the original and adversarial cases, which vary not only in the feature space but also in the semantic space, and then train an LSTM network to discover any discrepancy in the logit distribution in the semantic space. They offer a logit-based adversarial sample detection strategy that is very flexible, simple to implement in all pre-trained models, and has robust detection resistance against both black-box and white-box attacks.

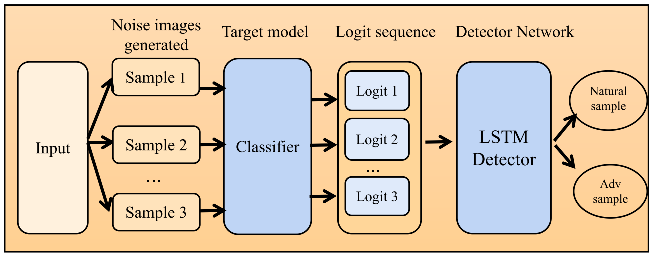


Fig. 15. An overview of the architecture of the logit-based adversarial sample detection method. The model analyzes the original and adversarial examples that differ not only in the feature space but also in the semantic space, and then trains an LSTM network to learn the differences in the logit distribution in the semantic space.

To detect arbitrary adversarial attacks without access to reference spectrographs and adversarial perturbations, Esmaeilpour et al. [225] propose a regularized logistic regression model to distinguish the eigenvalues of malicious spectral graphs from legitimate spectral graphs. They reveal that the manifolds of the adversarial samples are distant from the natural and noisy instances that are slightly disturbed by Gaussian noises. They use the eigenvalues of the legal examples and adversarial examples to train a logistic regression to find the decision boundary between them. This detector's main obstacle is its sensitivity to intra-class sample similarity, particularly in the multi-classification problem of black-box attacks.

Existing methods, e.g., [226], [227], focus on the visual field, and cannot detect adversarial examples in the radio signal field, which is an important domain due to the ubiquitous networks. In response to the adversarial attacks in the realm of radio signals, Xu et al. [58] describe a novel adversarial sample identification method by means of the integration of many features. They also provide a framework for creating adversarial samples, collecting local intrinsic dimension (LID) characteristics and constellation diagram (CD) characteristics, and recognizing adversarial samples. The framework produces the values of each layer for both normal and adversarial examples of the model. Then it computes the LID eigenvalues of the instance by estimating the maximum probability of a defined range of neighborhoods. The CD eigenvalues are computed simultaneously using the range characteristics and density features of the CD distribution. A logistic regression classifier is trained using several feature fusion values to identify adversarial samples. Experiments demonstrate that the suggested approach can reliably identify hostile radio signals. However, the performance degrades slightly when the perturbation is less

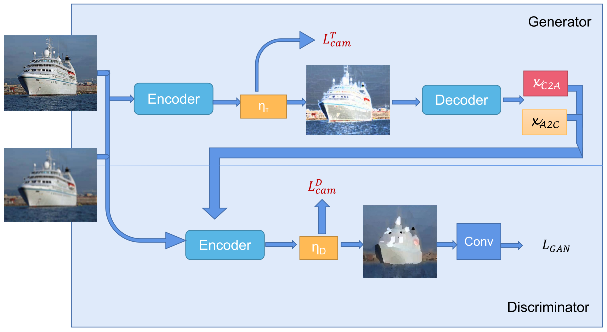


Fig. 16. Overview of the Generator and Discriminator in CAP-GAN. To accomplish adequate purification under cycle-consistent learning, the framework trains the purification model to increase robustness. It employs the standard GAN training approach to purify adversarial inputs by considering pixel-level and feature-level consistency.

than 10%. The reason is that the perturbations are very small, and the features are inconspicuous between the normal and adversarial examples.

All these methods, such as [218] and [58], represent the latest defense techniques that detect and defend against adversarial attacks. Adversarial sample detection is a typical defense method that aims to detect whether a sample is undergoing an adversarial attack directly. Logistic regression and deep learning are often used to classify adversarial and non-adversarial samples. On the one hand, adversarial sample detection methods can effectively detect and defend against adversarial attacks, in particular, black-box and white-box attacks. They are versatile and widely applicable to pre-trained models. On the other hand, adversarial sample detection methods may require large amounts of training data and high-capacity models with high computational overhead. They may have poor generalization capabilities and may be vulnerable to transfer attacks. Moreover, some of the methods are sensitive to intra-class sample similarity, especially in the multi-classification problem of black-box attacks.

2) Data Preprocessing: For defense methods for preprocessing input samples, classical research methods include PixelDefend [228], feature compression [229] and random transformations [230], etc. Table XI classifies the recent data preprocessing methods from the perspectives of attacked objects, input types, and invisible metrics.

Kang et al. [231] propose a purification model named Cycle-consistent attentional purification GAN (CAP-GAN), aimed at decreasing the impact of adversarial perturbations by transforming the input. The design of CAP-GAN is plotted in Fig. 16. The framework trains the purification model to enhance the pre-trained model's robustness in image classification. It uses the standard GAN training process to clean up adversarial inputs by balancing pixel-level and feature-level consistency via cycle-consistent learning for effective purification. On the CIFAR-10 dataset, CAP-GAN surpasses other preprocessing-based defenses, such as the JPG compression method developed in [232], in both black-box and white-box configurations.

To address the issue that preprocessing-based defenses are typically sensitive to the error amplification effect, Zhou et al. [233] suggest a self-supervised adversarial training mechanism in the class activation feature space to eliminate adversarial noise. Given adversarial assaults in the realm of radio signals, they first produce adversarial instances by substantially disrupting natural examples' class activation features. Then they train a denoising model, also known as the class activation feature-based denoiser (CAFD), by minimizing the disparities between the adversarial and normal examples in the class activation feature space. Consequently, antagonistic noise may be minimized. In comparison to prior approaches, such as the adversarial perturbation elimination GAN (APE-GAN) method developed in [234], the method developed in [233] considerably improves adversarial robustness, particularly against unexpected adversarial and adaptive attacks. However, for white-box attacks, the protection capability of the defense model is compromised as the defense model is completely visible to the attacker. In this sense, the defense against white-box adaptive attacks needs to be strengthened.

With the aim of addressing both adversarial examples deliberately created to cause harm and inputs that fall outside of the expected distribution, Wei and Liu [235] develop XEnsemble, a diversity ensemble verification technique. To limit the harm caused by malicious or incorrect data inputs, XEnsemble, an input-output model verification ensemble protection technique, may automatically examine every input to the prediction model. To protect DNN prediction models from adversarial examples and out-of-distribution inputs, XEnsemble uses a variety of data cleaning strategies, including rotation, color-depth reduction, local spatial smoothing and non-local spatial smoothing (NLM), to generate diverse input denoising verifiers, implements an ensemble learning approach to protect the DNN model from deception, and offers a set of algorithms for combining the input and output verifications. XEnsemble performs well in recognizing out-of-distribution inputs and protecting against adversarial samples. In order to make XEnsemble more resistant to internal attacks on the defensive system, the team has planned to randomize the input denoising integration layer and the output model validation layer, and also to generalize the XEnsemble technique to additional media types, including text, video, and audio.

By taking inspiration from the robustness domain [236], [237], [238], Zhu et al. [239] examine adversarial training from the standpoint of data-to-decision boundary distance. They introduce Saliency Adversarial Defense (SAD). This batch normalization strategy can achieve adversarial robustness without adversarial training by processing inputs through their saliency map and changing the Batch Normalization (BN) statistics. Compared to adversarial training, SDA is efficient in protecting against various forms of white-box and black-box attacks. It is generally anticipated that fine-tuning the processed data and adjusting the saliency map intensity based on the sample could lead to further improvement in performance.

When reducing adversarial noise, many existing model-agnostic defenses lose key image content, resulting in low classification accuracy on benign images. In this regard, Mustafa et al. [240] put forth an image restoration approach by using super-resolution, which projects off-the-manifold

TABLE X
ADVERSARIAL ATTACK DETECTION METHODS

Defense	Brief description	Similarity Measure	Evaluation	Strength	Weakness	Impacted Area
ROSA [218]	Shuffles pixels of an image and introduces some new universal noise to disrupt the adversarial perturbation then learns to predict the saliency mapping of the input image.	Energy function of low-level similarity	Precision, Recall, MAE, PR curves, F_β-measure	Significantly improves the backbone network robustness to adversarial attacks; shows comparable or better performance on natural images.	Only focuses on non-target attacks.	computer vision
CMAG [219]	Detects generative adversarial examples that are aware of the cascade model and demonstrates to humans what the model perceives by reconstructing the logit and feature maps of the final convolution layer.	SSIM	Detection Accuracy	Detects high-quality adversarial examples effectively and is more interpretable, providing a new perspective on the existence of the adversarial example.	The generative model is a simple autoencoder, whose performance is unsuitable for a complex dataset.	computer vision
ADNet [222]	A steganalysis-based deep learning model in which an attention module is incorporated to focus more on susceptible areas during feature learning.	ℓ_2 distance	Detection Accuracy	End-to-end, without manually extracting features.	Detection rates on normal images are lower than those on adversarial images created by C&W.	computer vision
Logit-based adversarial detection [224]	Trains an LSTM network to learn variations in logit distribution in semantic space and suggests a logit-based adversarial example identification technique.	logit	ROC, AUC	Detects the logit sequence differences between the original and adversarial examples, and is model-agnostic with strong generalizability.	Contrasting tests with other methods are missing.	computer vision
UnMask [223]	Based on robust feature alignment, detects adversarial perturbations by selecting a similarity threshold and safeguard the model by predicting an accurate class.	Jaccard Similarity	ROC, TP, FP	It is significantly more difficult for an attacker to modify all of the different characteristics that make up the image at the same time.	Only considers untargeted gray-box attacks.	computer vision
Regularized logistic regression model [225]	A regularized logistic regression model for discriminating eigenvalues of malicious spectrograms from benign ones.	Chordal Distance	AUC	Be able to detect any adversarial attack when no reference spectrogram or adversarial perturbation is available.	Significantly sensitive to intra-class sample similarities, especially for black-box multi-classification.	environmental sounds
LID&CD [58]	A detector based on multi-feature fusion, including generating, extracting LID&CD features and detecting adversary.	LID	Detection Accuracy	Innovatively detects adversarial examples in the radio signal domain instead of the vision domain.	When the perturbation is less than 10%, the performance is slightly decreased.	radio signals

adversarial instances into the native image manifold. The method is a model-independent defense mechanism that enhances an image by selectively adding high-frequency elements meanwhile canceling out any unwanted noise introduced by the attacker. Not only does the approach defend against attacks, but it enhances image quality while keeping model performance constant on clean images as well. Even when the attack model and attack type are unknown, the method performs better than white-box settings in terms of robustness.

Despite its importance in CNN prediction, the accurate recovery of input image structures has been generally overlooked in existing adversarial defensive systems. Yan et al. [188] develop D2Defend, which recovers both low and high-frequency picture structures in the spatial and transform domains, while eliminating adversary distortions. Unlike previous input-transformation approaches, such as a feature distillation method developed in [229], D2Defend uses bilateral and short-time Fourier transform (STFT) filtering to divide the input image into edge and texture feature layers. D2Defend is simple to develop and model-independent. It has been demonstrated to outperform existing adversarial defense approaches, particularly in high-attack scenarios. D2Defend's loss in clean accuracy is likewise judged acceptable and more stable than other defense methods.

These data processing-based methods [188], [189], [190], [191], [192], [193], [194], [195], [196], [197], [198], [199], [200], [201], [202], [203], [204], [205], [206], [207], [208], [209], [210], [211], [212], [213], [214], [215], [216], [217], [218], [219], [220], [221], [222], [223], [224], [225], [226], [227], [228], [229], [230], [231] can be broadly categorized as model-specific or model-agnostic defenses [241], and input-transformation or input-verification based defenses [242]. One commonality among these methods is that they aim to strengthen DNNs' resistance to adversarial attacks without significantly degrading their performance on non-attacked (clean) inputs. Some of the strengths of the methods discussed include their ability to defend against both white-box and black-box attacks, their model-agnosticism, and their ability to recover image structures of the input. Additionally, some of the methods are able to improve image quality and maintain model performance on clean images, and some of the methods are easy to deploy. On the other hand, some of the methods are found to be weak in defending against white-box adaptive

TABLE XI
DATA PREPROCESSING METHODS FOR ADVERSARIAL EXAMPLE DETECTION AND DEFENSES

Defense	Attack type	Brief description	Input	Invisibility Metric	Strength	Weakness
XEnsemble [235]	White-box, OOD	Improves DNN robustness against OOD inputs and adversarial examples by using diversity ensemble verification.	Image	ℓ_p	Automatically validates any input to the predictive model, be attack-agnostic. Superior in robustness and defensiveness, with a high defense rate of adversarial samples and a high detection rate of OOD inputs.	Needs to include randomization at the input denoising integration layer and the output model validation layer to increase resistance to internal attacks, as well as expand to new media.
SAD [239]	White-box, Black-box	Achieves adversarial robustness via analyzing inputs using saliency maps and updating BN statistics without AT.	Image	ℓ_2	Widens the average distance between the processed data and the updated decision boundary, significantly smooth the landscape, and is more effective than AT. Reduces the training time significantly, not relying on gradient masking.	Needs to fine-tune the processed data and adjusts the significance map intensity related to samples to further improve this method's performance.
Image super-resolution [240]	White-box, Black-box, Gray-box	A super-resolution-based image restoration technique that projects off-the-manifold adversarial samples into the benign image manifold.	Image	ℓ_1	No need for training or tuning many hyper-parameters. Do not cause gradient masking. Performs well for both black-box and white-box attacks. Supports unknown attacks.	Its robustness against white-box attacks is weaker than that in the gray-box settings.
CAP-GAN [231]	White-box, Black-box	Uses the pixel-level and feature-level consistency for GAN's cycle-consistent learning to achieve adequate purification.	Image	KL	The introduction of feature-level items fairly enhance model robustness. In both black-box and white-box conditions, CAP-GAN beats alternative preprocessing-based defenses on the CIFAR-10 dataset.	Adversarial interference is mitigated at the cost of removing important information from clean images, making DNN models less accurate for clean samples.
CAFD [233]	White-box, Unseen Attack	Devises a self-supervised AT technique in the class activation feature space to eliminate adversarial noise.	Image	CAFA	Compared to previous SOTA methods, the confrontation robustness is significantly enhanced, especially against unknown adversarial and adaptive attacks.	For white-box attacks, the defense model is completely leaked to the adversarial, and the protection capability of the defense model is destroyed. The defense against white-box adaptive attacks needs to be strengthened.
D2Defend [188]	White-box	Maintains the essential high-frequency image structure and filters out adversarial perturbations.	Image	SSIM	It is independent of DNN models and deploy-friendly. Good transferability among different commonly used networks and adversarial attack methods. The clean accuracy degradation is acceptable. More stable performance.	The defense effect under C&W attack is not optimal.

attacks, and others are found to cause gradient masking based on characteristics. Also, some of the methods proposed require fine-tuning the processed data and adjusting the significance map intensity related to the sample to improve the performance further.

3) Summary: Adversary detection approaches offer a strategy to defend against adversarial attacks in recent years. These methods attempt to detect adversarial samples in the input data and reject them, rather than modifying the original models and inputs. The advantages of adversarial detection include the ability to be used in combination with other defense methods and the ability to analyze whether the inputs contain an adversarial sample when the results of the baseline and robust classifiers do not agree.

On the other hand, adversarial detection methods have limitations. Some methods, such as LID&CD [58], ADNet [222], and UnMask [223], extract feature dimensions to detect adversarial samples. Other methods, such as CMAG [219], detect adversarial examples based on sample reconstruction comparisons. However, LID&CD [58] has insignificant detection performance when the perturbation is small. ADNet [222] has a better detection success rate for adversarial samples than for clean samples [225], and higher sensitivity for intra-class samples. They may also be bypassed by attackers who understand the detection mechanism. In the future, a combination of detection and defense is expected to be a promising direction to pursue.

C. Robustness Enhancement for Deep Learning Models

Current methods for defending against adversarial attacks focus primarily on improving model robustness. This goal is accomplished by incorporating regularizers into the model's loss function to make it more smooth. In other words, the gradient is Lipschitz continuous [243]. The goal is to make the model less sensitive to irrelevant variations in the input and off-manifold perturbations through effective regularization. The recent studies improving adversarial robustness can be broadly categorized into four main layers that regularization can be deployed: The input layer, middle layer, output layer, as well as across the layers.

TABLE XII
REGULARIZATION ON INPUT LAYER (I)

Defense	Strategy	Attack type	Description	Input	Inv-Metric	Strength	Weakness
AMM [247]	Add Noise	White-box, Black-box	A regularization method based on learning that uses an adversarial perturbation as a proxy.	Image	ℓ_p	Significantly improves the test set accuracy of various DNN architectures, and the generalization capability has been enhanced	Higher computational cost.
GAN-based Deepfake [248]	Add Noise	White-box, Black-box, Gray-box	An adversarial face generating approach that incorporates random differentiable image alterations during DeepFake model training to safeguard people's faces.	Image	ℓ_1	Makes it easy to recognize induced fake images and videos, regardless of model or data, and uses the same technique in all scenarios.	DeepFake model training takes a significant amount of time.
PDA [249]	Add Noise	Black-box, Corruption	During training, gradually introduces various adversarial perturbations.	Image	ℓ_2	Spends less training time while maintaining a high level of accuracy on clean samples, regulating the perturbation boundaries, ensuring greater robustness.	It is difficult to achieve robustness against white-box adversarial attacks that are confined to multiple spaces.
adMRL [143]	Add Noise	White-box	A novel MRL algorithm learning strategy for generalizing a meta-policy by meta-training an agent in a distorted environment with disturbed states.	Robotic trajectories	ℓ_p	Based on model-agnostic meta-learning, the agent may learn the initial parameters with improved generalization ability, as well as fight against additional "bad" samples.	Only tested on FSGM and random noise attacks.
HIRE-SNN [250]	Add Noise	White-box, Black-box	Uses the time steps of SNN training to efficiently input multiple noisy variations of the same image.	Image	ℓ_p	The robustness is not primarily derived from gradient masking, and the degradation in clean image accuracy is negligible. Improves inference latency and computation energy.	Improves model robustness at the expense of clean-image classification accuracy.
ICAT [251]	AT	White-box, Black-box	Adversarial training with one additional induced class.	Image	KL, CE	Proposes a new idea that the main impact of counteracting attacks is the alternation of prediction distributions.	Worse defense against black-box attacks over white-box attacks.
TriATNE [254]	AT	White-box, Black-box	A novel framework for learning stable and robust node embeddings with three participants, where the producer and the seller compete to win customers.	Node	ℓ_p	TriATNE outperforms the baseline on all datasets in link prediction, showing good performance on homogeneous networks.	Learning using, e.g., knowledge graphs, continues to pose difficulties.
Graph-AT [214]	AT	White-box	Dynamic regularization according to the graph structure.	Node	CE	Dynamic regularization; When learning to construct and resist perturbations, the influence of the connection instance is considered.	The computational overhead linearly rises as sampling more neighbors. Focuses only on graph-based learning from one graph.
CSA& CSE [258]	AL	White-box	CSA is a cost-sensitive adversarial learning approach. CSE is an end-to-end learning strategy that can improve the model's adversarial robustness with no need for AT.	Image	ℓ_2	Cost-sensitive training can successfully defend specific categories against adversarial attacks and can be used in conjunction with AT to increase performance.	May not be successful in improving the adversarial robustness of the model in a more complex dataset.
GAT [255]	AT	Unseen Attack, OOD	A strategy for improving the model's generalization to test data and OOD samples while also improving its robustness against unknown adversarial attacks.	Image	ℓ_p	Achieves SOTA robustness without training; improves both robustness and generalization. Extends to classification, semantic segmentation and object detection.	Maybe high computation cost.

1) Regularization on Input Layer: The essence of input layer regularization is to strengthen the generalization ability of neural networks through data enhancement. This can process input images, such as projection to non-adversarial manifold [244] and image conversion [245]. It can also train models by adding noise or using pseudo-labels in a semi-supervised learning mechanism [246]. Tables XII and XIII summarize the latest breakthroughs in the regularization of the input layer, divided into several aspects, including strategy, input, attack type, and inv-metric.

a) Noise perturbation: Noise Perturbation means processing input samples by injecting some mask or noise. Adversarial

TABLE XIII
REGULARIZATION ON INPUT LAYER (II)

Defense	Strategy	Attack type	Introduction	Input	Invisibility Metric	Strength	Weakness
APR [259]	AL	White-box, Black-box, Corruption, OOD	Recombines the phase spectrum of the current image with the distracter image amplitude spectrum to create a new training sample with the current image as the label.	Image	CE	Good adaptability to common corruption, surface changes, and OOD detection, while maintaining good ability on clean images.	Needs to investigate ways to represent part-whole hierarchies in neural networks based on the phase spectrum as well as more CNN models or convolution operations.
ART [252]	AL	Gaussian Noise	A neural network retraining strategy that indirectly improves a model's ability to maximize the least distance from all data examples to the decision.	Image and data points	ℓ_∞	The negative impact on classification accuracy is small. Reduces compute resources that spent on strengthening robust data samples and increment of model retraining time.	Experiments on small datasets only, which lack defensive performance experiments under strong attacks.
Style-Mix [256]	AL	White-box	A new data augmentation mix-up method that can generate different training samples by convex combinations of content and style characteristics.	Image	ℓ_2	Performs better or similar to SOTA mix-up approaches and learns more robust classifiers against adversarial attacks. Enhances the generalization of model training.	The automatic selection of the mixed ratio slightly degrades the defensive performance of further separate foreground from background.
MaxUp [253]	AL	Gaussian Noise	Introduces gradient-norm regularization for improving the loss function's smoothness to increase generalization performance and reduce overfitting.	Image, video, 3D point cloud	ℓ_2	MaxUp can easily take any SOTA data enhancement scheme and significantly improves them by minimizing the worst-case (rather than average) risk on enhanced data.	Although it merely necessitates an additional forward pass, MaxUp incurs a non-negligible additional time cost, which may be reduced using low-resolution images.
EdgeNet-Rob & Edge-GAN-Rob [260]	Feature Extract	White-box, Black-box, Corruption	Proposes two edge-enabled pipelines: EdgeNet-Rob and EdgeGAN-Rob. Makes CNNs rely more on edge features.	Image	CE	Edge features can improve the CNN model's robustness. Clean accuracy can be increased on datasets with clear edge information by repopulating the texture information.	Clean accuracy is slightly reduced.

Margin Maximization Networks (AMM) are provided by Yan et al. [247] as a learning-based regularization, which substitutes an adversarial perturbation for the geometric margin. By carefully crafting aggregation and shrinkage algorithms, AMM directly improves the classification margin. For many DNN designs, AMM greatly enhances test set precision while maintaining training set accuracy, demonstrating increased generalization power. Nevertheless, the computational cost of AMM increases due to the usage of repeated updates to estimate the classification margin and the calculation of high-order gradients during optimization.

By integrating random differentiable picture transformations when training DeepFake models [29], Yang et al. [248] present a transformation-aware adversarial face generation strategy to increase the defense capability against GAN-based DeepFake variants in the black-box situation. The technique can persistently yield more distortions in simulated face images, making it simpler to identify generated counterfeit images and videos, which are independent of models and data. The process follows the same steps under all settings. One potential downside is that the process of model training can be time-consuming.

Yu et al. [249] propose Progressive Diversified Augmentation (PDA), which increases the resilience of DNNs by gradually infusing different adversarial sounds in the training phase. This improves the system's overall resilience against adversarial examples and common corruption. Not only does PDA employ gradient information to make adversarial noises with negligible additional time cost, but it also uses a progressive schedule to vary the magnitudes of inputs during the training process. However, attaining robustness concurrently with PDA for white-box adversarial attacks restricted to numerous locations may be challenging due to the incompatibility of different adversarial robustness.

To address concerns pertaining to network security and network traffic analysis, McCarthy et al. [69] propose an innovative defense strategy, leveraging hierarchical learning to constrict the attack surface that adversarial examples may exploit given the constraints of an anticipated attack's parameter space. This robust defense learning model can withstand meticulously crafted adversarial attacks, maintaining classification accuracy on par with the original ML model when not under attack.

To enhance the robustness of Meta Reinforcement Learning (MRL), Chen et al. [143] propose adversarial Meta Reinforcement Learning (adMRL) to generate adversarial attack instances by using an adversarial GAN (adGAN) and leverage the generated examples to enhance the MRL algorithm's robustness. AdGAN and MRL can obtain good results by optimizing a minimax objective function during training. Building on model-agnostic meta-learning, the agents can learn the initial parameters with better generalization ability. Thus, when facing an unknown new task, the agents can learn to counteract these "bad" samples. However, the experimental attack methods only test on FSGM and random noise

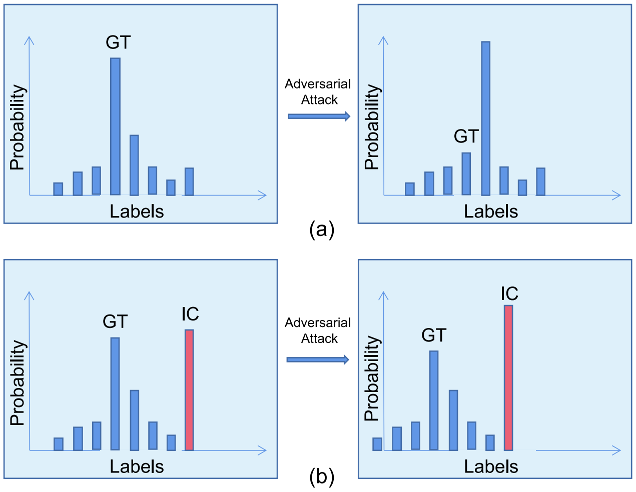


Fig. 17. An illustration of the alternation of the predicted distribution before and after an adversarial attack. (a) Traditional DNNs; and (b) The proposed induction method. Here, "GT" stands for "Ground Truth."

attacks, which may not be sufficient to fully demonstrate the robustness of the method.

Deep spiking neural networks (SNNs) modeled after the brain have gained popularity owing to their ability to reduce the power consumption of deep learning applications. In their study, Kundu et al. [250] present a spike-timing-dependent back-propagation (STDB)-based SNN training method to better leverage the inherent robustness of SNNs. Instead of continually displaying the same image, this approach makes use of the temporal phases of SNN training to input several noisy copies of the same image, hence decreasing the requirement for intermediate gradient storage and eliminating superfluous training time. The enhanced robustness of this method is not a result of gradient masking, and it demonstrates outstanding performance under black-box and white-box attacks, with a negligible loss in clean accuracy.

b) Adversarial learning: Adversarial learning aims to improve network security by improving the robustness of machine learning algorithms. As the main branch of adversarial learning, adversarial training, which involves training DNN models with adversarial examples, is another strategy for strengthening DNN robustness. Adversarial examples are produced with known adversarial attack algorithms, e.g., those described in Section III. Adversarial training can also be viewed as a special case of data augmentation that differs from traditional methods. Rather than introducing randomly transformed examples to improve model generality, adversarial learning introduces adversarially perturbed data to strengthen the model's robustness.

According to [251], the principal consequence of adversarial attacks is the modification of the prediction distribution. On the basis of this, they suggest Induced Class Adversarial Training (ICAT), a simple but successful strategy that incorporates an extra-induced class to defend against adversarial examples. Fig. 17 illustrates the alteration of predicted distributions before and after an adversarial attack. The method demonstrated better defense against white-box attacks compared to black-box attacks.

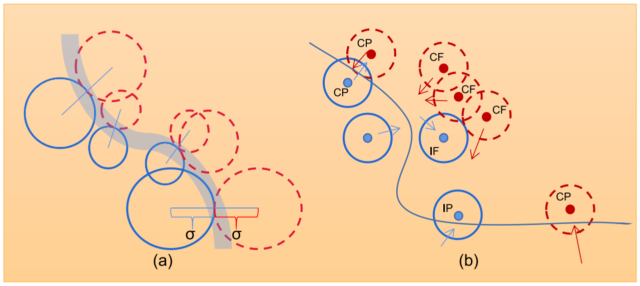


Fig. 18. The min-max game's and ART's concepts: Different classifications of data, denoted by color and circle style, compete for a larger influence area in the middle of the circles. (a) The gray area where the influence regions overlap should be the location of the optimum robust decision border. (b) The current decision boundary is shown by the solid line.

Yao et al. [252] target at classification and proposed Adaptive Retraining (ART) for neural networks, which implicitly improves a model's capacity in maximizing the minimal distance from data instances of all classes to the decision border. ART additionally builds a feedback loop and steers the data-generating process with categorization results for data augmentation. Fig. 18 shows the concept of the min-max game and ART. ART has a negligible negative impact on classification accuracy, reducing the computational resources for data augmentation and time increment for model retraining, making it suitable for the online optimization of neural networks. However, experiments have only been conducted on small datasets, lacking the verification of defensive performance on large-scale datasets under strong attacks.

Gong et al. [253] present MaxUp, a simple yet effective method for enhancing generalization and minimizing overfitting. The objective of the technique is to construct a collection of enhanced data with randomly generated perturbations or modifications, and then to minimize the greatest loss across the improved data. To improve generalization performance, the method implicitly incorporates a smoothness or robustness regularization against random disturbances. While an additional forward pass is all that is required, MaxUp still has non-negligible additional time costs that can be lowered by employing low-resolution picture acceleration during selection.

Tripartite Adversarial Training for Network Embeddings (TriATNE) is an adversarial learning system designed by Liu et al. [254], which learns stable and durable node embeddings with three players. A basic producer collects features for node pairings, a dynamic seller selects negative samples, and a biased consumer perturbs the objective function, all at the same time measuring the performance of node embeddings. Producing and selling are in competition for consumers. TriATNE is founded on the idea that a resilient approach must be able to endure interruptions and assaults. The TriATNE framework outperforms baselines on link prediction across all datasets and performs well on homogeneous networks, despite limitations in heterogeneous network learning.

Poursaeed et al. [255] propose Generative Adversarial Training (GAT) to boost the model's generalization ability to test sets and out-of-domain data, and its resilience against

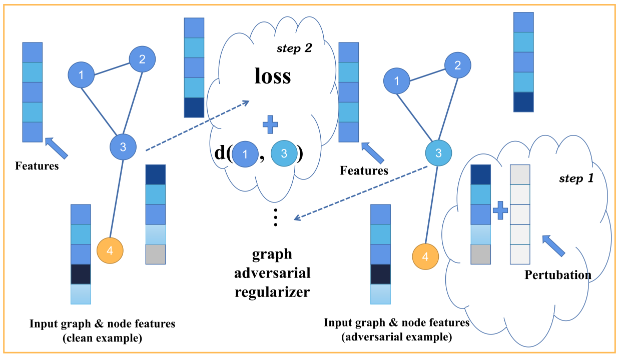

step 2
loss
d(1 , 3)
graph adversarial regularizer
Features
Features
step 1
Pertubation
Input graph & node features (clean example)
Input graph & node features (adversarial example)
1
2
3
4

Fig. 19. The training process of GraphAT: step 1) Producing graph adversarial instance and step 2) Optimizing model parameters through minimizing the loss and graph adversarial regularizer.

unanticipated adversarial attacks. GAT utilizes generative models with a disentangled latent space to produce a variety of low-, mid-, and high-level adjustments as opposed to modifying a single image feature. In addition to improving the model's performance on clean images and out-of-domain data, adversarial training with these cases makes it more robust to unforeseen attacks. The method is applicable to several applications, including classification, semantic segmentation, and object identification.

Hong et al. [256] claim that properly blending the content and style of two input images can result in more numerous and robust samples, which enhances model generalization during training. Based on this concept, they present StyleMix, a new data augmentation approach that provides a variety of training samples via convex combinations of content and style attributes. They further expand this technology to StyleCutMix, which enables sub-image level modification through CutMix's cut-and-paste methodology [257]. They also devise a method for automatically determining the degree of style mixing based on the class distance between two images. Experiments show that their strategies increase classifier robustness against adversarial attacks more than other recent mixup methods and improve model training generalization.

Feng et al. [214] propose a dynamic regularization scheme called GraphAT, as shown in Fig. 19. The scheme breaks the smoothness of linked nodes to the greatest degree possible and generates network adversarial examples by perturbing the input of associated clean examples. Furthermore, it minimizes the graph neural network's objective function using an extra regularizer across adversarial graph samples. This promotes smoothness between adversarial and linked example predictions, making the model more robust to perturbations transmitted across the graph. However, the computing cost grows linearly with the neighbors sampled. In addition, the work focuses only on graph-based learning from a single graph, but future research objectives seek to determine the efficacy of graph-adversarial training on numerous graphs.

Shen et al. [258] suggest a novel method for safeguarding a particular class from adversarial attacks, which is different from previous defensive approaches that try to increase the resilience of overall classes. They adopt CSA and cost-sensitive adversarial extension (CSE) to include cost sensitivity in adversarial learning and enhance the model's adversarial robustness. Fig. 20 depicts an overview of the adversarial learning system, as well as the CSA and CSE algorithms. The techniques have been tested on the MNIST and CIFAR datasets. However, the gain in robustness for more complicated datasets may be limited.

In view of human outstanding generalization ability, Chen et al. [259] argue that a resilient CNN should be able to endure changes in amplitude while concentrating on the phase spectrum. In order to do this, they provide a unique data enhancement approach dubbed Amplitude-Phase Recombination. APR integrates the phase spectrum of the current picture with the amplitude spectrum of an adversarial image to generate a new training sample with the same label as the current sample. This strategy allows the CNN to get more structured data from the phase components than from the amplitude components. APR is at the forefront of several generalization and calibration problems, such as adaption for surface fluctuations and common corruptions, adversarial assaults, and out-of-distribution detection.

c) Feature Extraction: To train a model, feature extraction focuses on the texture features of images. Sun et al. [260] are particularly interested in shape features and propose two edge-enabled pipelines, namely, EdgeNetRob and EdgeGANRob, to force CNNs to rely more on edge features, inspired by the fact that the visual system of humans ends up paying more attention on global features, such as shapes, for recognition, whereas CNN models are biased towards local features (e.g., textures) in images. Both EdgeNetRob and EdgeGANRob use an edge detection technique to extract structural attributes from a picture. EdgeNetRob then trains downstream learning tasks using the recovered edge features, while EdgeGANRob rebuilds a new image by filling in texture features using a learned GAN. These findings demonstrate adding edge features can increase the model's robustness while decreasing clean accuracy significantly.

2) Regularization on Middle Layer: Intermediate layer regularization can be achieved by operating on neurons, hidden layers, and Lipschitz condition constraints. Tables XIV and XV briefly introduce the latest regularization methods on the middle layer and compare their advantages and disadvantages.

Layer regularization: On the middle layer, the most commonly adopted regularization method is layer regularization, where a regularization term is added to the hidden layer.

Adversarial Noise Propagation (ANP) [261] is a simple yet strong training technique that, unlike classic adversarial defensive methods, does not manipulate solely the input layer (as discussed in Section IV-C1). During training, it injects noises into the hidden layers by propagating backward from the adversarial loss. This enables the learned parameters in each layer to produce accurate and consistent results for the benign instance and its distributed noisy surrogates, resulting in a high degree of resilience. Due to the fact that each layer helps improve the resilience of the model to differing degrees in this technique, it is necessary to build a more adaptable

TABLE XIV
ROBUSTNESS REGULARIZATION ON MIDDLE LAYER (I)

Defense	Strategy	Attack type	Description	Input	Invisibility Metric	Strength	Weakness
KR [276]	Lipschitz	White-box	A new optimal transport-based classification framework that incorporates the Lipschitz constant and the gradient norm preservation constraint as a theoretical need.	Image	hinge-KR	The expected guarantee in terms of robustness is supplied without any major loss of accuracy. Uses adversarial attacks to interpret a prediction.	The calculation time increases during learning.
Inf-Norm& Inf-Ind [215]	Lipschitz	White-box	Two novel robust training approaches for CNN architectures are proposed. Both techniques require modifying the network structure using an approximate non-smooth regularization term that influences the spectral constants in the convolutional layers and fully-connected layers.	Image	ℓ_2, ℓ_∞	Improves the architecture's robustness by making network mapping more reliable and interpretable. Gradually increasing complexity through warm-starting.	The use of non-smooth terms restricts the application of traditional gradient-based learning techniques.
ANP [261]	Layer Reg.	White-box, Black-box	A powerful training algorithm that adds noises to the hidden neural network layers in a layer-wise manner.	Image	RMS Distance	Easily integrated with adversarial training methods, and efficiently performs using the basic backward-forward approach, introducing no additional computations and memory consumption.	Since each layer contributes to the robustness of the model to varying degrees, a more adaptive algorithm needs to be designed to take into account the heterogeneous behavior of the different layers.
SNS [263]	Layer Reg.	White-box, Black-box, Gaussian Noise	Stabilizes neuron sensitivities toward benign and adversarial examples.	Image	ℓ_1	SNS performs well against black-box attacks on different datasets, such as SPSA and NATTack, but does not achieve security through obfuscation.	Updating sensitive neurons dynamically incurs higher computational costs.
HLDR [264]	Layer Reg.	White-box	A training technique and goal function for arbitrary neural network designs that are adversarially robust.	Image	ℓ_1	More effective than other advanced techniques to protect neural networks from adversarial evasion attacks; only requires as many parameters as the original model; training fairly quickly.	The effect of distortion, different training methods, and attack strategies on HLDR performance is unclear.
ER-Classifier [266]	Layer Reg.	White-box, Black-box	A new end-to-end robust DNN defensive scheme that enhances the classifier's adversarial robustness by embedding regularization.	Image	WD	In addition to improving the classification accuracy of adversarial samples, the framework can also be used to detect adversarial samples.	Further research on low-dimensional space needs to be done to increase the robustness of DNNs.
RP [267]	Reg.	White-box	Proposes RP-Ensemble, a training technique consisting of projected versions of the original inputs and RP-Regularizer, a regularization term to the training objective.	Image	ℓ_1	Totally independent of the attack's selected norm, and computationally efficient.	On CIFAR-10, the results are less significant than those for MNIST.
DH-AT [265]	Reg.	White-box	An upgraded variation of AT that connects a second head to one of the network's intermediate layers.	Image	CE, KL	Be readily added with minor changes into existing adversarial training techniques, and can achieve clean precision and robustness at the same time.	Incurs more training time.

algorithm that takes into consideration the changing behavior of different levels. While perturbing the deep layers boosts the model's robustness, adding noises to the shallow layers has an adverse effect. Nevertheless, the root source of this issue is not studied in [261].

Regarding countermeasures to adversarial attacks in the context of communication networks, researchers have made significant strides in adversarial defense. Dong et al. [262] put forth an innovative defense mechanism, underpinned by a GAN framework. The proposed model sheds light on the role of adversarial attacks and defense within an end-to-end learning process of communication systems, utilizing an approach comprising triple training. Specifically, it involves the joint adversarial training of encoder and decoder communication neural networks against adversarial attacks. Zhang et al. [2] propose a defense system against adversarial samples in transformer-based modulation classification, aiming to transfer the trained adversarial attention mapping from a large transformer to a more compact transformer. This contributes to robustness in the face of adversarial attacks. They used a special adversarial attack, i.e., the white-box PGD algorithm, to generate adversarial examples. It is proved that the transformer-based neural network is more robust to PGD attacks than CNN.

TABLE XV
ROBUSTNESS REGULARIZATION ON MIDDLE LAYER (II)

Defense	Strategy	Attack type	Description	Input	Invisibility Metric	Strength	Weakness
TENET [270]	Layer Reg.	White-box	A regularization strategy based on group-wise inhibition for increasing feature diversity and network robustness.	Image	CE	Enables the network to explore varied and richer features, resulting in a more accurate picture representation, even when malicious alterations are introduced.	Maybe high computation cost.
SACNet [271]	Layer Reg.	White-box	A new self-attention context network in which all of the loss at one pixel is shared by all of the pixels that are connected to it. Attacking such a network requires a higher level of perturbation.	Hyper-spectral Image	Context enhanced features	The HSI's global contextual information may considerably enhance the resilience of DNNs against adversarial attacks.	SACNet has a higher time cost since self-attention learning and context encoding raise the computing overhead of the whole framework.
FNC [176]	Layer Reg.	White-box	Feature norm-clipping layers, which are differentiable modules that may be arbitrarily placed in different CNNs, are used to prevent the creation of excessively large norm deep feature vectors.	Image	ℓ_p	Does not introduce trainable parameters and has only very low computational overhead, effectively improving the robustness of white-box generic patch attacks by different CNNs.	The identification accuracy of the clean samples affects slightly. The reason why remaining valid for the location-independent patch for a single image is unknown.
Wave-CNets [272]	Layer Reg.	White-box, Black-box	For noise-resistant image classification, DWT is applied to the feature maps during downsampling to minimize aliasing.	Image	Corruption Error	Separation of aliasing effects, or the differentiation between low- and high-frequency information, may be stymied without resorting to adversarial training.	It is less resistant to attackers than specially trained defense methods, and the wavelet transform introduces additional computations.
Supervision Layer [25]	Layer Reg.	Black-box, Gray-box, White-box	A defensive model in which a supervisory layer is introduced as an auxiliary classifier to the basic neural network model.	Image	ℓ_p	The quality of the features recovered by the hidden layer is enhanced by the black-box and gray-box threat models.	In the case of white-box defense, it will only reduce the confidence of the attacker and make them unable to protect against assaults effectively.
Adv-Rush [273]	Layer Reg.	White-box, Black-box	A unique NAS adversarial robustness-aware neural architecture search technique that employs a regularization term generated from the neural network's loss landscape curvature data.	Image	ℓ_2	AdvRush successfully discovers a neural process for adversarial robustness, which is highly transferable on different datasets.	For large regularization strength, the searched architecture experiences a significant drop in clean accuracy.
Cross-Domain Ensembles [274]	Layer Reg.	White-box, Corruption	A general framework for making robust predictions based on creating a diverse ensemble of various middle domains.	Image	ℓ_1	No manual modification or redesign is required when making a change between middle domains. No additional supervision or labeling is required than what the dataset already comes with; Can be extended to a whole new non-adversary and anti-corruption.	Depends on reasonable uncertainty estimation in the presence of distribution transfer; Selection of intermediate domain; Using only unimodal distribution. Integration-based methods can increase computational complexity.

Zhang et al. [263] provide a unique perspective of neuron sensitivity to explain adversarial resilience for deep models, as assessed by the magnitude of variance in neuronal activity in response to benign and adversarial situations. They suggest a Sensitive Neuron Stabilizing (SNS) approach after analyzing the behaviors of the model's intermediate layers and demonstrating dependence between adversarial resilience and neuron sensitivities. By stabilizing sensitive neurons (instead of obfuscation), the technique tries to increase the model's robustness to adversarial samples. However, dynamically updating sensitive neurons incurs a greater computational expense.

Yao and Gao [25] offer a defensive model that makes use of a supervision method to enhance the model's robustness. The assumption is that supervision can increase the quality of feature maps for the hidden layer, hence increasing the resilience of the model. As a secondary classifier, a supervisory layer is appended to the main neural network model to perform this strategy. By utilizing the classification results of intermediate layers to assess perturbation properties and by continuously changing the loss function of the supervisory layer, the quality of the hidden layer's recovered features may be enhanced. This decreases the impact of adversarial perturbation and boosts the model's overall resilience. Under both black-box and gray-box threat models, the suggested technique can survive assaults from FGSM as well as C&W, BIM, and DeepFool attack algorithms to a large degree. However, under a white-box threat model, it will only reduce the confidence of the attacker and cannot effectively defend against attacks, being considered a weakly supervised defense model.

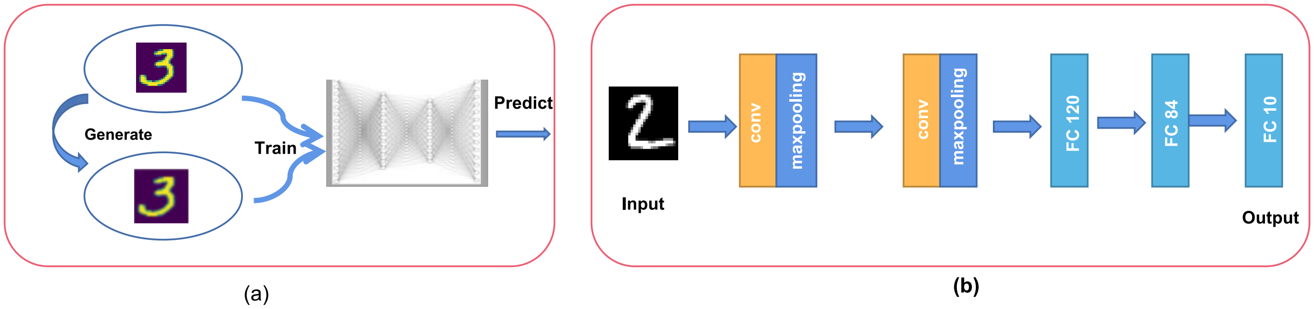


Fig. 20. Comparison of conventional adversarial learning and CSA/CSE algorithms. (a) In adversarial learning, clean and adversarial examples are alternately supplied into the neural network. (b) To accomplish the min-max property, the framework for CSA and CSE algorithms applies the convolution parameter to the loss function by adding a normalization.

Taking steps to shield ML-based radio signal (or modulation) classification from adversarial attacks, Zhang et al. [59] investigate a defense mechanism based on train-time and run-time defense techniques. The train-time defense consists of adversarial training and label smoothing, while the run-time defense employs neural rejection (NR) based on support vector machines (SVMs). Specifically, the system uses label smoothing and Gaussian noise augmentation (LS-GNA) while adopting a stronger form of attack generated by customized Adversarial Training (AT) to generate adversarial perturbations for each sample according to the parameters and architecture of the CNN. The disturbance level and corresponding label are customized in the process of adversarial training.

Schwartz and Ditzler [264] propose a method called HLDR to improve adversarial robustness by using latent disparity regularization. The regularizer is defined to depend linearly on the disparity in representations created in the hidden layers based on benign and adversarial data. The regularizer penalizes the discrepancy and provides significant improvements in adversarial robustness while also reducing training time. However, the method only considers training programs with fine-tuned subsets constructed using one method. It is not yet understood how HLDR performance may be affected by adversarial, Gaussian or other forms of distortion, different training methods for fine-tuning subsets, and different objective functions.

Dual Head Adversarial Training (DH-AT) [265] is a novel defensive method that employs a dual-headed architecture to increase both clean accuracy and adversarial resilience as an enhanced form of Adversarial Training (AT) in both network structure and training strategy. The architecture of the lightweight CNN used by DH-AT is outlined in Fig. 21. As seen in the diagram, DH-AT connects a second network head (or branch) to a network's intermediate layer before combining the outputs of both heads using a lightweight CNN. In order to attain both clean precision and resilience in the meantime, the training technique is modified to account for the relative significance of the two heads. A potential drawback is that DH-AT may require longer training time.

Zheng et al. [47] introduce PUAA with the aim of assessing the robustness of DNN-based spectrum sensing models. The PUAA method introduces a meticulously engineered disturbance to the benign primary user signal, resulting in a substantial reduction in the detection probability of the spectrum sensing model. In response, they propose an autoencoder-based defense method named DeepFilter to counteract PUAA, as shown in Fig. 22. The integration of LSTM neural networks and CNNs within DeepFilter enables simultaneous extraction of temporal features and local features of the input signal, contributing to its effective defense capabilities. Experimental evidence validates that DeepFilter can efficiently guard against PUAA, without compromising DNN-based spectrum sensing mode's detection performance.

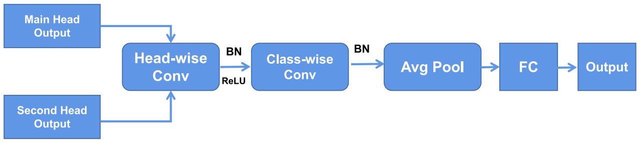


Fig. 21. The workflow of DH-AT. DH-AT connects another head to an intermediate layer of the neural network, and utilizes a shallow CNN to integrate the outputs of the two heads.

Li et al. [266] propose an Embedding Regularized Classifier (ER-Classifier) to improve the adversarial resilience of classifiers. The intrinsic dimension of image data is substantially less than its pixel space dimension, and hostile examples often dwell outside the manifold of natural image data. The approach projects high-dimensional input images into a low-dimensional space and returns adversarial samples to the manifold of natural image data via regularization. This enhances classification accuracy when faced with adversarial scenarios. In addition, the framework may be utilized in conjunction with detection approaches to discover adversarial instances. Exploration of low-dimensional areas to improve the resiliency of DNNs is a potential innovation.

Similarly, Carbone et al. [267] propose RP-Ensemble, a training approach based on the Manifold Hypothesis [268], [269]. Using projected representations of the original inputs, this method enhances the robustness of a pre-trained classifier against adversarial examples. Moreover, they develop the RP-Regularizer, a regularization term based on the norms of the loss gradients, which measures vulnerability, with the expectation over random projections of the inputs. This is done during training to capitalize on pertinent adversarial characteristics. The strategy is computationally efficient and independent of the attack norm type. However, the CIFAR-10 dataset produces less impressive results than the MNIST dataset.

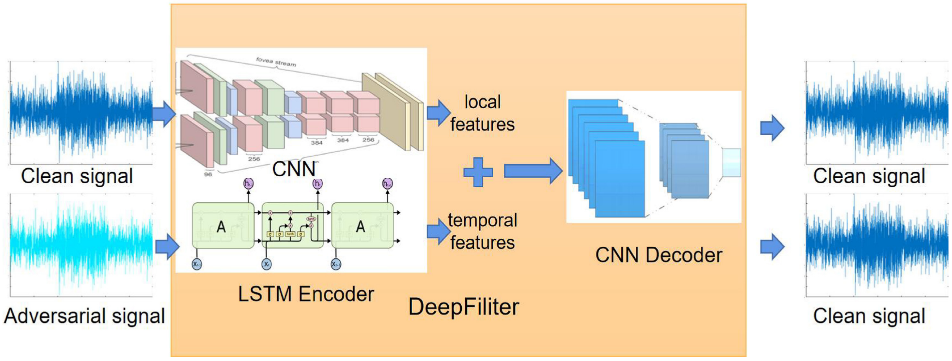


Fig. 22. The defense framework based on DeepFilter, which is composed of encoder and decoder. The encoder extracts the local and primary features, and the decoder reconstructs the signal according to the extracted features. DeepFilter does not affect the benign signals and can process the adversarial signals as benign signals.

Boora et al. [66] conduct an in-depth investigation into the implications of adversarial attack and defense mechanisms, specifically adversarial training, on massive MIMO localization utilizing CNN and ODE model. The authors propose a novel Neural ODE method, combining convolution blocks, ODE blocks, and dense layers to achieve a localized regression solution. The adversarial training of this Neural ODE helps to resist adversarial samples, thereby improving the robustness of massive MIMO localization.

CNNs usually ignore key auxiliary properties, whilst current adversarial training and regularization approaches overlook the independence of local features. Liu et al. [270] introduce TENET Training, a group-wise inhibition-based regularization approach for enhancing feature diversity and network resilience. The suggested approach dynamically regularizes CNNs during learning by suppressing regions with the highest activation values that are the most discriminative. This allows the network to study more diverse aspects, which can more accurately depict pictures, even when they have been altered maliciously. TENET Training improves both robustness and generalization significantly in comparison to other SOTA methods.

While classifying hyperspectral pictures, a novel self-attention context network (SACNet) [271] is presented to strengthen the network's inherent resilience to adversarial samples. Existing adversarial learning algorithms are designed to recognize RGB pictures, but this strategy targets hyperspectral images (HSIs). In contrast to local feature extraction, global context information extraction necessitates the construction of associations between a specific pixel and all relevant pixels in the whole picture. This pixel's network prediction would be impacted by its neighboring pixels, and the overall loss would be dispersed across all neighboring pixels. Therefore, tackling these networks may need a greater degree of disturbance. The suggested SACNet has a very high temporal cost since self-attention learning and contextual encoding raise the computing burden of the overall framework. Future research should also examine if SACNet can defend against adaptive RGB images.

According to [272], the downsampling method is primarily responsible for CNNs' poor noise resilience. They combine frequently employed CNN designs with a discrete wavelet transform (DWT) to produce wavelet-integrated convolutional networks (WaveCNets), which address the issue of aliasing in CNNs and enhance noise resistance in image classification. During downsampling, DWT separates the WaveCNets feature maps into low-frequency and high-frequency components. Low-frequency components are passed to succeeding layers to retain robust high-level characteristics, while high-frequency components are deleted to avoid noise transmission. Although WaveCNets regularly resists different forms of noise, its performance is inferior to that of well-trained defensive systems. The wavelet transform also adds computational complexity.

Yu et al. [176] show that universal adversarial patches in prevalent CNNs often generate deep feature vectors with large norms. They suggest a simple but effective defensive technique based on a unique feature norm clipping (FNC) layer. This differentiable module can be dynamically introduced to various CNNs to prevent the adaptive creation of huge norm-deep feature vectors. An FNC investigation of the effective receptive field (ERF) reveals that the adversarial patch's impacts may be minimized naturally, resulting in enhanced classification accuracy against adversarial patch attacks. Experiments conducted on multiple datasets demonstrate that this proposed technique enhances the robustness of various CNNs against white-box universal patch assaults while retaining acceptable identification accuracy for clean samples and incurring a little computational cost. However, it remains unclear why this method is still valid for position-independent image patches.

Mok et al. [273] investigate the topic of constructing an adversarially resilient neural network with strong inherent resilience and robust training strategies. Adversarially robust architecture rush (AdvRush) is an adversarial robustness-aware neural architecture search (NAS) approach based on the observation that the inherent robustness of a neural network depends on the smoothness of its input landscape regardless of the training procedure. Using a regularizer that prefers candidate architectures with a smoother input loss landscape, AdvRush selects a neural network that is resistant to adversarial inputs. The approach is very adaptable to many datasets. Future studies will examine the robustness of neural network architecture on multimodal datasets and broaden the search area to include activation functions.

Yeo et al. [274] present a general framework for developing robust predictions based on the creation of a varied ensemble of different middle domains. The suggested method makes predictions by combining a variety of cues (called "middle domains") into a single strong prediction. The concept is that predictions based on distinct cues react differently to a distribution adjustment. As a result, they can be combined into a robust final prediction. The method can change without manual modification or redesign, or any additional supervision or labeling already attached to the dataset. It can generalize to new non-adversarial and anti-corruption scenarios. On the other hand, the limitations of the method include its reliance on reasonable uncertainty estimates in the presence of distribution shifts and its use of only unimodal distributions for the study. Additionally, the selection of middle domains and the use of ensemble-based methods inevitably increase computational complexity.

b) Lipschiz condition: To mitigate the sensitivity of the output to changes in the input, Amini and Ghaemmaghami [215] offer a new non-smooth regularization term in the optimization formulation and two non-smooth regularizers that construct direct linkages for the weight matrices in each neural network layer. These regularizers adjust the Lipschitz constant of the underlying architecture to make the mapping between inputs and outputs more stable, hence reducing the network's sensitivity to small perturbations. However, regular gradient-based learning methods become less useful when non-smooth variables are included.

It has been demonstrated in [275] that limiting a neural network's Lipschitz constant gives certifiable robustness guarantees against local adversarial attacks and increases the model's generalization ability and interpretability. Therefore, Serrurier et al. [276] provide a novel optimal transport-based classification framework that takes into account the Lipschitz constant and the gradient norm preservation requirement. They use a regularized Kantorovich-Rubinstein formulation that includes a hinge loss term, which provides the desired robustness guarantees with little accuracy loss. One possible drawback is that the computation time increases during learning.

Some defensive approaches augment the standard training aim for intermediate layers with graduated penalties. For example, by using a layer-by-layer contrast penalty term, Gu and Rigazio [277] are able to retain the output of a DNN unaffected by input disturbances. FNC [176] is also a novel feature norm shear layer that can be placed flexibly into various networks to adaptively suppress the creation of big norm depth eigenvectors and enhance its overall performance.

3) Regularization on Output Layer: The output layer can be regularized through defense distillation, label smoothing, or by modifying loss functions. An overview of the possible regularization methods in the output layer is provided and compared in Table XVI. As delineated in Section III, adversarial attacks can be primarily categorized into three types depending on the nature of the attacks: black-box attacks, white-box attacks, and gray-box attacks. White-box attacks, despite being the most potent variant of adversarial attacks, are arguably the least prevalent in practical applications. They are frequently employed to assess the robustness of defense and/or classification models under stringent conditions. In the case of black-box attacks, the attacker is devoid of any knowledge concerning the internal structure, training parameters, and defensive mechanisms of the targeted model, and can only engage with the model through its output. Consequently, defense methods predicated on gradient masking face challenges in resisting black-box attacks. During a gray-box attack, the attacker has access to the classification model but lacks any information about its defense strategy. Gray-box attacks represent a middle ground and can be very useful for evaluating the robustness of defenses and classifiers.

a) Distillation: Many recent works enhance the resilience of a student network with a teacher network by using Knowledge distillation (KD) [278] in combination with adversarial training. They build on adversarial training by using a teacher network that has already been pre-trained in the form of adversaries. According to [187], adversarial training approaches are more successful on large models and less effective on small models. To solve this issue, Zi et al. [187] propose RSLAD, a unique method for training small, robust student models by distilling from adversarially trained large models. This strategy replaces hard labels with robust soft labels inside supervision loss terms. It employs the huge, robust instructor model's robust soft labels to assist student learning in both natural and adversarial situations. However, one downside of the strategy is that when the teacher network grows too complicated for the student network to learn from, the student's robustness tends to decline.

Attention Guided Knowledge Distillation and Bi-directional Metric Learning (AGKD-BML) is a new adversarial training-based model proposed by Wang et al. [279]. By leveraging knowledge distillation, the approach is made up of two parts: Bidirectional attack metric learning (BML) and attention-guided knowledge distillation (AGKD). To enhance the attention map for adversarial instances and repair damaged intermediate features, the AGKD module extracts clean image attention information to the student model. The BML component employs bidirectional metric learning to standardize the feature space representation. The combination of these two modules consistently surpasses cutting-edge techniques, including single-directional metric learning (SML) [280], Bilateral [281], and feature scattering (FS) [282]. The authors further demonstrate that the model's robustness is not generated from gradient obfuscation, but rather from a slight drop in clean accuracy.

Beamforming prediction is integral to the advancement of next-generation wireless networks. In this regard, Kuzlu et al. [283] highlight the security vulnerabilities associated with employing DNN for beamforming prediction in 6G wireless networks. They portray the prediction as a multi-output regression problem and offer two mitigation methods - iterative adversarial training and defensive distillation methods. These strategies successfully enhance the predictive performance of RF beamforming, generating more accurate predictions. Additionally, the proposed scheme demonstrates efficacy even when adversarial samples contaminate the training data. Experimental results substantiate that the method can

TABLE XVI
ROBUSTNESS REGULARIZATION ON OUTPUT LAYER

Defense	Strategy	Attack type	Description	Input	Invisibility Metric	Strength	Weakness
AGKD-BML [279]	Distillation	White-box, Black-box	Attention Guided Knowledge Distillation and Bi-directional Metric Learning are used to create an adversarial training-based model.	Image	ℓ_2	By integrating AGKD and BML, achieves SOTA robustness under different attacks. Robustness does not come from gradient obfuscation.	AGKD-BML trained on a 7-step attack achieves much lower performance against the regular attacks. The clean accuracy has decreased a little.
Distilled Differentiator [284]	Distillation	White-box, Black-box	Activation-based network pruning is used to maximize the transferability of attacks and retrain their precision.	Image	ℓ_p	Effectively defends against targeted and non-targeted attacks, maintaining scalability, effectiveness, and comparable clean accuracy through a small-scale ensemble model.	Ensemble-based methods inevitably increase computational complexity.
RSLAD [187]	Distillation	White-box, Black-box	A unique adversarial robustness distillation approach for training robust small student models, whereby robust soft labels are substituted for hard labels in all supervision loss terms.	Image	KL	Improves the robustness of small models against SOTA attacks, especially automated attacks; more effectively than previous adversarial training and distillation techniques.	When the teacher's network grows too complex for the student to understand, the student's resilience declines.
PCL [287]	Loss	White-box, Black-box	A proactive defense against adversarial assaults, a novel distance-based training technique that aims to maximally segregate the learned feature representations at different depth layers of the deep model.	Image	PCL	Greatly enhances the robustness of the learning model, even against the strongest white-box attacks, without clean accuracy decline, not due to obfuscated gradients, and requires a shorter training time.	Robustness against white-Box settings is higher than black-box settings.
Luring effect [186]	Loss	Black-box	A novel strategy to thwart transferability of black-box attack between two models.	Image	CE	Luring effect can be implemented at a low cost for any pre-trained model, and can successfully limit the efficiency of even the most advanced transfer-based black-box attacks with large adversarial perturbations, and be effectively combined with existing defense schemes.	The advances in robustness are more evident on SVHN and MNIST, but the findings on CIFAR-10 are especially encouraging in the context of a defensive system that needs a pre-trained model.
BLF [289]	Loss	White-box, Black-box	The insertion of a new limited function shortly prior to softmax improves adversarial robustness.	Image	ℓ_∞	BLF is easily combined with AT, proving that BLF is superior to logits squeezing without AT, and is superior to or comparable to logit compression, label smoothing, and TRADES when AT is used.	Logit regularization methods without AT are insufficiently resistant to targeted assaults. It is unclear why small logit can improve robustness.
PER [290]	Bound	White-box	Promotes bigger adversarial-free zones in the vicinity of the input data, hence enhancing the proven robustness of the models.	Image	ℓ_p	Suitable for different architectures and networks with generic activation functions; the computational overhead of PER is negligible; achieves better robustness guarantees and accuracy for clean data.	Regardless of the linearization approach, the boundary of the output logarithm for large models unavoidably grows looser over deeper networks. Furthermore, the linear approximation implicitly prefers the ℓ_∞ norm over the ℓ_p norm.

effectively shield DNN models from adversarial attacks in the context of next-generation wireless networks.

Wu et al. [284] build a distilled differentiator using activation-based network pruning to decrease attack transferability, meanwhile retaining accuracy. As a two-phase defense, they use an ensemble structure of diverse differentiators. In the first step, the student model is utilized to narrow down the possible differentiators to be developed. In the second stage, a small, predetermined number of differentiators are employed to properly evaluate clean or reject hostile inputs. This solution fits the criteria for defense rate, model accuracy, and scalability. Through small-scale integration models, the architecture retains scalability, efficacy, and comparable clean input accuracy while being more efficient and simpler to implement. However, ensemble-based methods, such as boosting [285] and bagging [286], would inevitably increase the computational complexity compared to non-ensemble models, due to their need for extensive model training and a combination of multiple prediction results into a final result.

b) Loss function: Many training techniques have been developed to enhance performance by modifying or adding new regularization terms to the models' loss functions.

In addressing the modulation classification problem based on DNN, with the objective of crafting a DNN model resilient to attacks, Manoj et al. [63] introduce three defense techniques: random smoothing, Hybrid Projection Gradient Descent (HPGD) adversarial training, and rapid adversarial training. These methods have been assessed under the

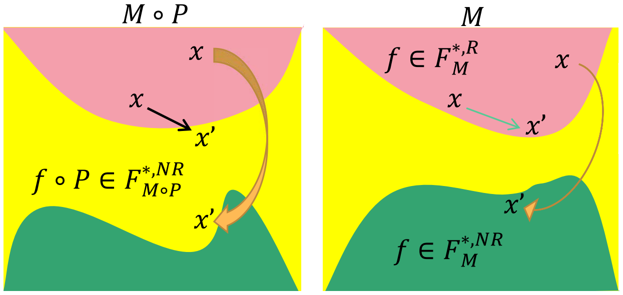


Fig. 23. On the left, x' dupes $M \circ P$ by flipping the non-robust feature $f \circ P$ (to the green class). On the right, however, f can be a robust feature, in which case M will not be deceived (still in the pink class).

conditions of both white-box and black-box attacks. The findings reveal that rapid adversarial training exhibits superior robustness and computational efficiency compared to other techniques, and it is capable of generating models demonstrating robust defenses against realistic attacks.

Mustafa et al. [287] explain the proximity of distinct classes of samples in the learned feature space of deep learning models is the primary reason for the vulnerabilities in DNNs. As a proactive protection against adversarial attacks, they suggest a distance-based training technique, Prototype Conformity Loss (PCL), to solve this issue. This strategy aims to maximally segregate the learned feature representations at many layers of a DNN so that there is little intersection between any two classes in the decision layer, as well as the intermediate feature space. Such a strategy assures that an adversarial instance with a limited perturbation budget can no longer deceive the network. The strategy significantly increases the model's robustness with a shorter training period, without diminishing the classification accuracy of clean pictures. However, the model is less resistant to white-box attacks than it is to black-box attacks since gradient masking is not utilized in its defense.

Bernhard et al. [186] create an innovative technique known as the "luring effect" to prevent transferability between two models to pave a new path for robustness in a realistic black-box situation. They provide a deception-based method that is applicable to any pre-trained model and needs no labeled dataset. The target model is reinforced with a neural network designed to have an appealing effect and trained with a loss function that employs logit sequence order.

Fig. 23 illustrates the two cases of the Luring effect. On the left, x' dupes $M \circ P$ by flipping the non-robust feature $f \circ P$. However, on the right of the figure, f can be a robust feature, in which case M will not be deceived (still in the pink class), or a non-robust feature, in which case $f \circ P$ toggle will not be tracked. Even with massive adversarial perturbations, the approach can successfully limit the efficiency of cutting-edge transfer black-box attacks. It may be effectively coupled with existing defense strategies. The benefits of the method for robustness are more pronounced on the SVHN and MNIST datasets, but the CIFAR-10 results only perform well within the range of defense schemes that need a pre-trained model.

Like conventional method for improving adversarial resilience by limiting logit norms to tiny values [288], Kanai et al. [289] introduce a function named bounded logit function (BLF), which employs a bounded activation function shortly before the softmax to confine the logit norms. BLF is constrained by finite values. Moreover, its pre-logit at the maximum or minimum points is constrained by finite values. Consequently, introducing BLF right prior to softmax gives finite values to the optimal logit and pre-logit. Despite its simplicity, the approach successfully enhances robustness in adversarial training compared to alternative logit regularization methods. However, although BLF is effective against powerful non-target attacks, it is useless against target attacks. Furthermore, the approach reveals the softmax cross-entropy's fragility, as well as the efficiency of logit regularization. However, it is unclear why a small logit enhances robustness.

c) Boundary Estimation: Liu et al. [290] concentrate on constructing certifiers to identify certified areas of the input neighborhood where the model produces the right prediction and employing such certifiers to train a model to be verifiably resilient to adversarial attacks. They developed a stronger certifier, the polyhedral envelope certifier (PEC), as well as a regularization scheme named polyhedral envelope regularization (PER), which can be applied to networks of different architectures with general activation functions. Different from some earlier methods, such as COnvex Layer-wise adversarial Training (COLT) developed in [291], PER has minimum computing cost and offers improved robustness guarantees and precision on clean data in a variety of circumstances. However, the method can be restrictive for large models, particularly for deeper networks. The boundary of the output logarithm inevitably becomes less tight, irrespective of the linearization method being used. In addition, the linear approximation is in favor of the ℓ_∞ norm over other ℓ_p norms, performing better in the case of ℓ_∞ than it performs in the case of ℓ_2.

4) Regularization Across Layers – Gradient Masking: Gradient masking here refers to a typical approach for protecting against white-box attacks that depend on model gradients. This method includes adding an extra training layer to the model, which decreases its sensitivity to tiny changes in the input data. This is often accomplished by using random noise or perturbations to obfuscate gradient information, or by employing a gradient near or at zero to neutralize or mitigate gradient-based attacks. However, gradient masking does not alter the decision boundaries. Instead, it just makes it more challenging for an attacker to influence the model using gradient information. This means that gradient masking is generally not effective against black-box attacks. The attacker can simply self-train an agent model to mimic the defense model, e.g., by observing the real discriminant labels of the input samples.

Many defense approaches have been developed based on gradient masking [292]. Some are directly designed to perform gradient masking, such as replacing the smooth sigmoid function with a hard threshold [293]. Some strategies add regularization terms with a gradient penalty component, making the model less susceptible to tiny input perturbations [214], [264]. However, this strategy has the potential to significantly diminish the model's precision and learning capability. Defensive distillation substitutes the last layer with a soft maximum function and a temperature junction to regulate the degree of distillation after the training process [284].

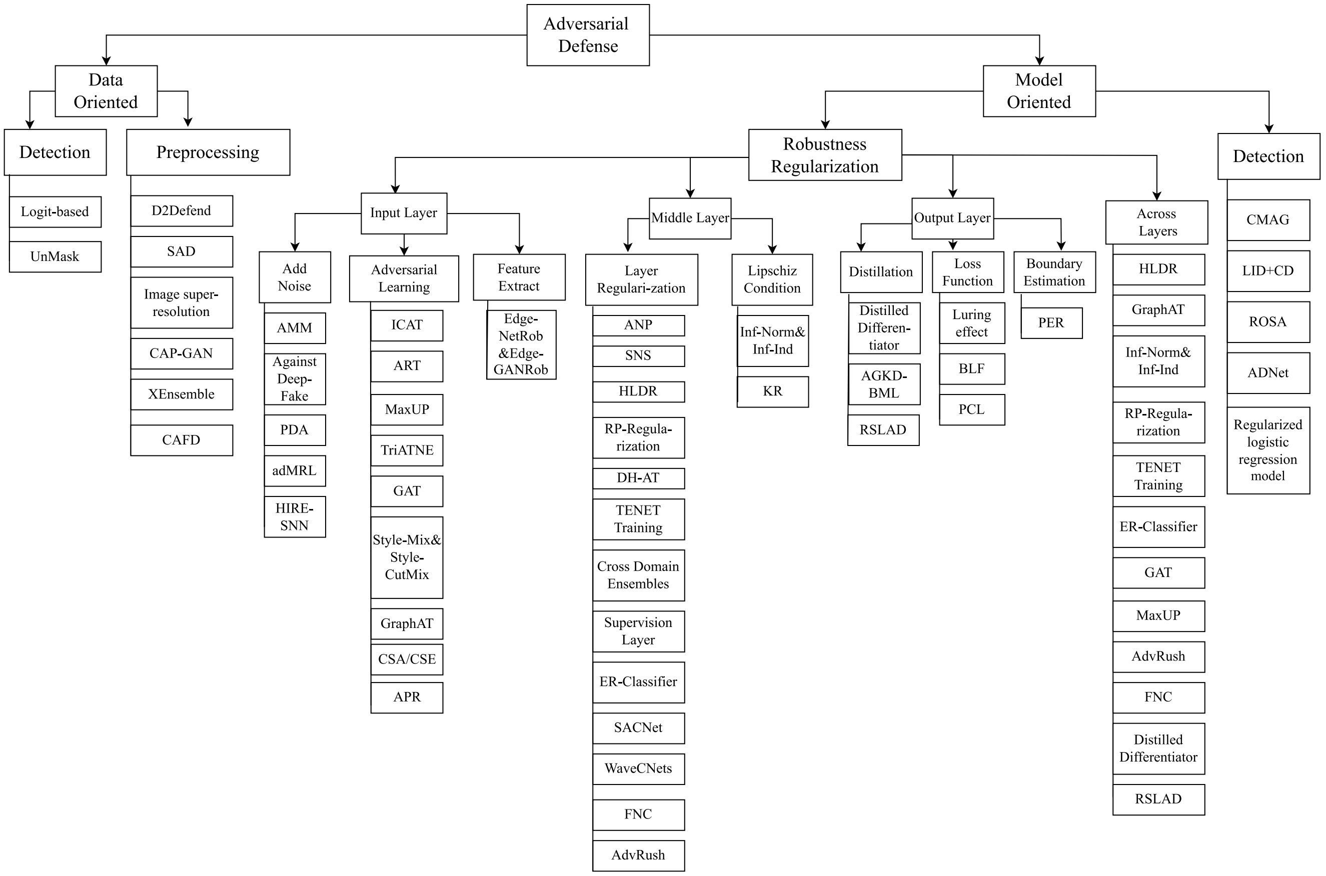


Fig. 24. Anatomy of breakthroughs in adversarial defenses since 2021. We classify adversarial defense methods as data-oriented and model-oriented methods. From the model-oriented angle, we further divide defense methods into robustness regularization in different layers and detection.

Gradient masking has used regularizers or smoothing labels to make the model less susceptible to input perturbations. Some of these strategies include blurring or masking gradient data, akin to gradient masking. HLDR [264], for instance, adds a regularization term to penalize the difference between benign and adversarial data representations in the hidden layer. GraphAT [214] aims to minimize an adversarial graph regularizer and reduce prediction divergence between a disturbed target instance and its related instances. Robust CNN Training with Inf-Norm and Inf-Ind regularization [215] is used to improve the total Lipschitz constant and consistency of input-output maps. However, these methods do not use a gradient step to update the penalty parameter, which can lead to ambiguity in the adversary. In other words, they also blur the gradient. The RP-Regularizer [267] integrates the specification of the loss gradient, intended as a metric of vulnerability, with the expectations of stochastic predictions of the inputs. To prevent steep gradients caused by binary masks, in TENET Training [270], researchers propose Rectified Reverse Function (RRF) to smooth the group inversion mapping. MaxUp [253] adds a gradient-norm smooth regularization for Gaussian perturbations. The regularization procedure in ER-Classifier [266] may assist in eliminating adversarial distortion effects and returning adversarial instances to normal data manifolds. GAT [255] provides variety and realism to adversarial training examples to close the distributional gap between adversarial and actual samples. AdvRush [273] introduces regularizers for candidate architectures that smooth input loss landscapes.

Defensive distillation is an additional kind of gradient masking [294], which substitutes the last layer with a protective soft maximum function and a temperature junction to modify the level of distillation after the training process. RSLAD [187] is a strategy that uses resilient, soft labels created by an adversarially-trained teacher model to guide the training of students on both clean and adversarial samples. Distilled Differentiator [284] utilizes an ensemble structure built on specialized classifiers called differentiators and activation-based network pruning to limit attack transferability while maintaining precision.

5) Summary: In the past few years, researchers have made many contributions to the field of regularization methods for adversarial defense from four perspectives: Input layer, middle layer, output layer, and across layers. Regularization methods at the input layer can be divided into noise addition, adversarial learning, and feature extraction. Amongst these, adversarial learning, which is the most effective and widely used defense method, has developed many variants in recent years. Regularization methods at the middle layer focus on changing the model structure to improve its inherent robustness. The output layer regularization can be classified as distillation, decision boundary estimation, and loss functions design, and it effectively improves the robustness of the models from different perspectives. Last but not least, certain defense techniques attempt to combat adversarial attacks across neural network layers by implementing gradient masking, which is ineffective in the face of black-box attacks.

D. Summary and Lessons Learned

The latest adversarial defense techniques, especially those published in the past few years, have been primarily focused on adversarial detection and robustness enhancement techniques. Adversarial detection methods are simpler and more effective than modifying the original model and input images, but they can be easily bypassed by attackers. On the other hand, robustness enhancement techniques aim to improve the accuracy of the model and reduce the success rate of attacks. However, there are still challenges, such as reducing data dependency, avoiding gradient masking effects, improving model generalization, reducing the cost of model training, and improving resistance to unknown high-intensity attacks.

Combining detection and robustness-enhancing defense methods is a promising direction for future research. Existing methods, such as GAT [255], XEnsemble [235], APR [259], and ER-Classifier [266], have already contributed to this perspective. Future research is anticipated to focus on developing methods that combine the advantages of both detection and robustness enhancement, while addressing the challenges mentioned above. This will help to improve the overall effectiveness of adversarial defenses.

V. Lessons Learned and Open Issues

In this section, we summarize the important lessons learned from the comprehensive analysis of the recent research outcomes in adversarial attacks and defenses, devise the remaining open challenges and point to research opportunities in this rapidly growing, important area.

A. Lessons and Challenges of Adversarial Attacks

It is important to strike a balance between effectiveness, imperceptibility, complexity, and transferability in adversarial attacks, among which there are obvious trade-offs. As illustrated in Fig. 25,

- Gradient-based attack methods, such as LAFEAT [125], DSNGD [127], SGA. [134], are known for their high ASRs and good transferability. However, the methods have limitations of high computational and time costs, as well as the issue of "gradient saturation," which reduces their effectiveness. Gradient-based attacks limit the perturbation to a certain size during the generation of the perturbation, guaranteeing invisibility.
- Constrained optimization-based attack methods, such as GF-Attack [149], AdMRL [143], and SSAH [147], have good transferability. However, they are known for high computational and time costs, making it difficult to use them in time-sensitive applications. Constraint optimization-based attack methods can guarantee small visibility of the attack by constraining the strength of the perturbation, providing greater stealth.
- Search-based attack methods, such as FeaCP [154], are highly transferable and can be extended to other domains beyond image classification. Nevertheless, for more complex datasets, such as ImageNet [252], searching for the optimal adversarial sample needs significantly more iterations and more computational cost. It can be difficult to find an appropriate search start point for search [295]. Current search-based methods are primarily applied to the initialization or optimization of other adversarial sample generation algorithms, such as Square attack [23]. The search-based perturbation does not make use of gradient information, and the perturbation magnitude of each search step is controlled to fall under a fixed range to ensure a certain invisibility.

Future efforts are expected to reduce attack costs, improve the transferability of attacks across different datasets and models, and extend the attacks to more deep-learning tasks.

The exploration of adversarial attacks within the field of communication networks holds significant value. Adversarial attacks have the capacity to dismantle meticulously trained network traffic classification models and deceive ML-based IDS employed within the IIoT [6]. This underscores the need for defensive solutions to mitigate these adversarial attacks within IoT networks.

Countermeasures against adversarial attacks can aid in addressing issues related to covert communication and privacy disclosure in communications [91], [296], [297]. For instance, in the context of RISs in wireless communication systems, adversarial disturbance and RIS interaction vectors can be co-designed to effectively enhance signal detection accuracy at the receiver end [298], [299]. Concurrently, this reduces the detection accuracy at the eavesdropping terminal, enabling covert communication [3]. Furthermore, by developing obfuscation methodologies for traffic types, malicious traffic type analysis (TTA) tactics can be misdirected, resulting in incorrect classification of traffic type or user activity, thus facilitating privacy protection [79].

B. Lessons and Challenges of Defenses

It is crucial to provide effective and reliable countermeasures to adversarial attacks for an apparent reason. Existing adversarial defenses can benefit from continuing development to address the following challenges.

1) Trade-Off Between Defense Effectiveness and Overhead: Designing a model that is robust against adversarial attacks is an important aspect of ML. Adversarial attacks attempt to mislead a model into making a mistake by slightly modifying the input data. They exploit the vulnerabilities of the model to achieve their purpose. Although mitigating these attacks is necessary for the reliability and security of ML models, it comes with its own overhead and challenges:

1) *Computational Overhead:* Enhancing the robustness of models often demands additional computational resources. Techniques, such as adversarial training, in which the model is trained on adversarial examples, are computationally expensive [54]. The requirement of adversarial training to repeatedly run adversarial attack algorithms for obtaining new adversarial examples against the optimized ML or DNN model within

TABLE XVII
LESSONS LEARNED AND OPEN ISSUES

Field	Lessons Learned	Open Issues
Attack	• Gradient-based attack methods: High ASRs and good transferability, but high computational and time costs, as well as the issue of "gradient saturation". • Constrained optimization-based attack: Good transferability but high computational and time costs. • Search-based attack methods: Highly transferable but difficult to find an appropriate search start point.	• Reducing attack costs. • Improving transferability across different datasets and models. • Extending to more deep learning tasks.
Defense	• Trade-off between defense effectiveness and cost: Training the model in a complex environment or structural improvements greatly boosts robustness with a cost of additional computational complexity. • Unanswered Root Cause of Robustness Loss: For bounding the Lipschitz Continuity of the gradient of DNNs, including Gradient Clip, Weight Decay, and gradient masking, which is controversial in the case of adversarial defenses. • Scalability and Generalizability: It is generally hard to apply a defense method that is effective on one DNN model or dataset, to other complex DNN models or datasets. • Data-driven: Detection methods for adversarial samples are more data-driven, but current research on detection techniques is limited by the lack of consensus on adversarial samples at the mathematical level.	• Developing methods that combine both detection and robustness enhancement. • Addressing challenges such as reducing data dependency, avoiding gradient masking effects, improving model generalization, reducing the cost of model training, and improving resistance to unknown high-intensity attacks. • Overall improving the effectiveness of adversarial defense techniques. • A combination of detection and defense approaches is the research trend.

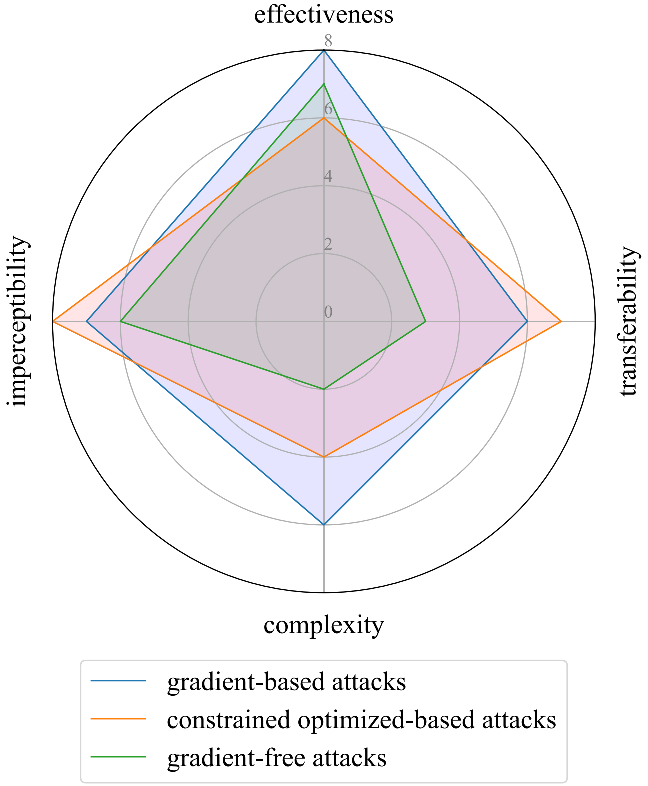


Fig. 25. The balance of four factors in adversarial attacks.

each training epoch further exacerbates this issue. This "optimal" adversary is typically procured via multi-step gradient descent, leading to a substantially extended time frame for model learning when using standard adversarial training methods, compared to conventional training techniques. Moreover, for large datasets and complex models, these methods can dramatically increase training times.

For example, adMRL [143] enables agents to learn the initial parameters with better generalization ability by accumulating knowledge in difficult environments, such that agents can learn the ability to fight these "bad" samples when faced with an unseen new task. PDA [249] uses different magnitudes of adversarial samples to increase the diversity of data, progressively injecting diverse adversarial perturbations during training, which, however, also increases the training cost. AMM [247] focuses on the minimum (instance-specific) margin, which is often considered a key factor in determining a model's generalization capability. A principled regularizer is derived to improve the model's performance on unseen samples with certain types of distortions. However, the requirements of training time and memory space are greatly increased as a result of iterative updating processes and the use of higher-order gradients in the optimization process.

2) *Increased Model Complexity:* Many methods of increasing model robustness involve adding complexity to the model architecture. This can involve augmenting the model with additional layers or nodes [25], increasing the risk of overfitting and requiring more data for effective training. For instance, Zhang et al. [300] use Graph Convolutional Network (GCN) and neural random forest to build an end-to-end learning system, in which the GCN module uses user information and evaluation information to capture user hobby information. The random forest module is used to detect malicious users. However, the model becomes complex by containing unusually many fully connected layers. The increased complexity can also lead to higher memory requirements and longer inference times, which can impact the overall performance of the model. Moreover, the upgraded model in DH-AT [265] contains two heads for robustness and clean accuracy independently, at the expense of higher training time. The framework for classification based on optimum transport in [276] contains a Kantorovich-Rubinstein (KR) regularization approach

and more accurate constant evaluation of convolution and pooling layers, but almost triples the training time. Because self-attentional learning and context coding raise the computing overhead of the overall framework, SACNet [271] also has a somewhat lengthy execution time. Due to the employment of numerous models, several ensemble-based methods [267], [284] generally increase inference complexity.

There have been many noteworthy efforts to reduce training costs, including transfer learning, partial training/updating, optimizing training epochs, and parallel training. For example, an adaptive retraining process is used in ART [252]. The Luring effect [186] can be used to improve a trained model at a low cost, since it does not require labeled datasets. It is also possible to introduce additional regularization items, such as HLDR [264], or layers, such as FNC [176], to neural networks for performance improvements without increasing model parameters. In harnessing the inherent robustness of energy-efficient deep spiking neural networks (HIRE-SNN) [250], the weight updating only takes place after T steps, thereby reducing the training cost while allowing different adversarial image variants to train the model. The iterative approach is only used for the unfrozen layer for the Distilled Differentiator [284]. PDA [249] selects the iterative steps that best balance robustness, precision, and computing cost. To parallelize training, random projections ensemble (RP-Ensemble) [267] creates classifiers in separately projected subspaces. In addition, models may be trained concurrently in Distilled Differentiator [284]. These efforts alleviate the trade-off between defense efficacy and expense to some degree, but they do not eliminate the trade-off.

It is worth noting that research has shown that there is often a trade-off between model accuracy on clean (non-adversarial) data and robustness to adversarial attacks. More robust models often have reduced performance on clean data [301]. Moreover, new types of adversarial attacks are also developed as new defense mechanisms are created. It is a continuous arms race, which adds to the overhead as models must be constantly updated and retrained to counter new attacks.

2) Unanswered Root Cause of Robustness Loss: Lipschitz continuity is a mathematical concept used to measure the degree of continuity between two functions [302]. More specifically, a function f from $S \subset \mathbb{R}^n$ into $\mathbb{R}^m$ is Lipschitz continuous at $x \in S$ if there exists such a constant C that, for all $y \in S$ close to x,

$$\|f(y) - f(x)\| \leq C\|y - x\|.$$

In the context of deep learning, Lipschitz continuity is used to measure the robustness of DNNs by assessing how small changes in the inputs affect the outputs of the DNNs. In other words, a Lipschitz-continuous function has a fixed-ratio bound on the distances between the corresponding outputs of two points close to each other in the input space [302]. A DNN is said to be Lipschitz-continuous if small changes in its inputs can only cause small changes in its outputs, implicating the stability of a DNN in the face of noisy data or unexpected inputs.

There are several methods for bounding the Lipschitz Continuity of the gradient of DNNs. A few common methods are Gradient Clip [303], which involves clipping the gradients to ensure they do not exceed a threshold, and Weight Decay [304], which involves adding a regularization term to the loss function of a DNN to limit the magnitude of the gradients.

Another popular method for bounding the Lipschitz Continuity of the gradient of DNNs is gradient masking [292], which involves adding a penalty term to the loss function to prevent abrupt changes in gradients. However, gradient masking is controversial in the case of adversarial defenses [292]. Some researchers have argued that although gradient masking-based defense techniques, e.g., ER-Classifier [266] and Inf-Norm&Inf-Ind [215], deliver effective defense against adversarial attacks in some cases, they do not address the root cause of adversarial attacks [292], which is the lack of Lipschitz Continuity. As a consequence, gradient masking approaches are vulnerable to attacks that are independent of the gradients of the models under attack, such as high-intensity black-box attacks [292].

3) Scalability and Generalizability: To defend against (new) adversarial attacks, one approach is to design more effective and robust neural network models [22], and the other is to block malicious inputs before it is fed into the model [305]. Since most heuristic defense strategies are unable to defend against adaptive white-box attacks, many researchers have begun to focus on provable defense mechanisms that guarantee a certain level of defense performance, irrespective of the attacker's method of attack [254].

Scalability has been a key issue to the majority of existing provable defense approaches, e.g., the PGD method developed in [22]. Proof-based defense strategies are effective in defending against less sophisticated "shadow" neural networks, but are ineffective in the case of more advanced "deep" neural networks [306]. Moreover, while provable defense methods work satisfactorily on small-scale datasets, e.g., the CIFAR-10 dataset with only ten classes, they deteriorate on more difficult tasks, such as classification on the ImageNet dataset that consists of a thousand major classes [252].

Generalizability, also known as transferability, is another major concern of provable defense approaches. Specifically, it is generally hard to apply a defense method that is effective on one DNN model or dataset, to other DNN models or datasets [267]. One approach may yield satisfactory results on homogeneous networks, but performs poorly on heterogeneous networks [254].

4) Dependence on Data: Adversarial sample detection has traditionally relied on data-driven methods. However, there is a lack of agreement on the mathematical definition of adversarial samples, limiting current research in this field [307]. Attackers can easily bypass detections by exploiting the knowledge of the detection mechanisms, rendering the detection mechanisms ineffective [308]. To overcome this challenge, a mixed approach that integrates both detection and defense strategies can offer a promising solution. The defense component aims to enhance accuracy and decrease attack success rate, while reducing data dependence and increasing resilience to high-intensity attacks. These are important research areas in pursuit of robust defense mechanisms.

VI. Conclusion

We have provided a comprehensive overview of the recent advancements in adversarial attacks and defenses in ML and DNNs with an emphasis on applications to communications and networking. We have analyzed both the attack techniques, including those based on constrained optimization and gradient-based optimization, and their adaptations to different threat models, such as white-box, gray-box, and black-box attacks. We have reviewed the latest defense strategies against adversarial examples, including detection and robustness improvement, which mainly focus on enhancing robustness through regularization, data augmentation, and structure optimization. Moreover, the transferability of adversarial attacks has been thoroughly investigated, providing deeper insights into the workings of DL models.

Our research highlights the significant impact of adversarial attacks on communication and networking. We have discovered that adversarial attacks can exploit vulnerabilities in ML or DNN-based functions, including wireless signal classification, modulation scheme recognition, and resource allocation in MIMO networks. We have also identified adversarial attacks in network management and NIDS, particularly in DNN-based traffic classification. While some initial defensive strategies have been proposed to combat the adversarial attacks, continuing efforts are required to address surges of new adversarial attacks that can potentially compromise communication and networking systems.

References

[1] Y. Zhao, W. Yuan, Y. Huang, and Z. Chen, "Adversarial attacks in DNN-based modulation recognition: A preliminary study," in *Proc. 9th Int. Conf. Depend. Syst. Appl.*, 2022, pp. 1061–1062.

[2] L. Zhang, S. Lambotharan, and G. Zheng, "Adversarial learning in transformer based neural network in radio signal classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 9032–9036.

[3] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Channel effects on surrogate models of adversarial attacks against wireless signal classifiers," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[4] K. S. Durbha and S. Amuru, "AutoML models for wireless signals classification and their effectiveness against adversarial attacks," in *Proc. 14th Int. Conf. Commun. Syst. Netw.*, 2022, pp. 265–269.

[5] C. Zhang, X. Costa-Pérez, and P. Patras, "Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms," *IEEE/ACM Trans. Netw.*, vol. 30, no. 3, pp. 1294–1311, Jun. 2022.

[6] S. Li, J. Wang, Y. Wang, G. Zhou, and Y. Zhao, "EIFDAA: Evaluation of an IDS with function-discarding adversarial attacks in the IIoT," *Heliyon*, vol. 9, no. 2, 2023, Art. no. e13520.

[7] T. Altaf, X. Wang, W. Ni, R. P. Liu, and R. Braun, "NE-GConv: A lightweight node edge graph convolutional network for intrusion detection," *Comput. Security*, vol. 130, pp. 103285–103294, Jul. 2023.

[8] Q. He, J. Liu, and Z. Huang, "WSRC: Weakly supervised faster RCNN toward accurate traffic object detection," *IEEE Access*, vol. 11, pp. 1445–1455, 2023.

[9] B. Kaur and S. Singh, "Object detection using deep learning: A review," in *Proc. Int. Conf. Data Sci. Mach. Learn. Artif. Intell.*, Aug. 2021, pp. 328–334.

[10] X. Yuan, S. Hu, W. Ni, X. Wang, and A. Jamalipour, "Deep reinforcement learning-driven reconfigurable intelligent surface-assisted radio surveillance with a fixed-wing UAV," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4546–4560, 2023.

[11] S. Hu, X. Yuan, W. Ni, X. Wang, and A. Jamalipour, "RIS-assisted jamming rejection and path planning for UAV-borne IoT platform: A new deep reinforcement learning framework," *IEEE Internet Things J.*, early access, Jun. 7, 2023, doi: 10.1109/JIOT.2023.3283502.

[12] C. Xu, W. Gao, T. Li, N. Bai, G. Li, and Y. Zhang, "Teacher-student collaborative knowledge distillation for image classification," *Appl. Intell.*, vol. 53, no. 2, pp. 1997–2009, 2023.

[13] L. Liu, "Improved image classification accuracy by convolutional neural networks," in *Proc. Int. Conf. Inf. Technol. Electron. Eng.*, Changde, China, Oct. 2021, pp. 1–5.

[14] Y. Wang, M. Zhao, S. Li, X. Yuan, and W. Ni, "Dispersed pixel perturbation-based imperceptible backdoor trigger for image classifier models," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3091–3106, 2022.

[15] G. Angelova, E. Avramidis, and S. Möller, "Using neural machine translation methods for sign language translation," in *Proc. Annu. Meet. Assoc. Comput. Linguist. Stud. Res. Workshop*, May 2022, pp. 273–284.

[16] T. Ananthanarayana et al., "Deep learning methods for sign language translation," *ACM Trans. Accessible Comput.*, vol. 14, no. 4, pp. 1–30, 2021.

[17] J. Y. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: A review of sentiment analysis based on sequential transfer learning," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 749–780, 2023.

[18] Z. Sun, L. Tian, Q. Du, and J. A. Bhutto, "Sample hardness guided softmax loss for face recognition," *Appl. Intell.*, vol. 53, no. 3, pp. 2640–2655, 2023.

[19] T. Abdullah and A. Ahmet, "Deep learning in sentiment analysis: Recent architectures," *ACM Comput. Surveys*, vol. 55, no. 8, pp. 1–37, 2023.

[20] A. Kelkar and C. Dick, "NVIDIA aerial GPU hosted AI-on-5G," in *Proc. IEEE 5G World Forum*, Montreal, QC, Canada, Oct. 2021, pp. 64–69.

[21] Y. Hu et al., "Artificial intelligence security: Threats and counter measures," *ACM Comput. Surveys*, vol. 55, no. 20, pp. 1–36, 2021.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Rep.*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–23.

[23] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 484–501.

[24] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy*, Jose, CA, USA, May 2017, pp. 39–57.

[25] Z. Yao and J. Gao, "Adversarial example defense based on the supervision," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[26] R. Duan et al., "Adversarial laser beam: Effective physical-world attack to DNNs in a blink," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 16062–16071.

[27] L. Xu, X. Zheng, X. Li, Y. Zhang, L. Liu, and H. Ma, "WiCAM: Imperceptible adversarial attack on deep learning based WiFi sensing," in *Proc. Annu. IEEE Int. Conf. Sens. Commun. Netw.*, Stockholm, Sweden, Sep. 2022, pp. 10–18.

[28] D. Chew, D. Barcklow, C. Baumgart, and A. B. Cooper, "Adversarial attacks on deep-learning RF classification in spectrum monitoring with imperfect bandwidth estimation," in *Proc. IEEE Wireless Commu. Netw. Conf.*, 2022, pp. 1152–1157.

[29] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surveys*, vol. 54, no. 1, pp. 1–7, 2022.

[30] S. Tariq, S. Jeon, and S. S. Woo, "Am I a real or fake celebrity? Evaluating face recognition and verification APIs under deepfake impersonation attack," in *Proc. World Wide Web*, Apr. 2022, pp. 512–523.

[31] L. Qin, F. Peng, M. Long, R. Ramachandra, and C. Busch, "Vulnerabilities of unattended face verification systems to facial components-based presentation attacks: An empirical study," *ACM Trans. Privacy Security*, vol. 25, no. 1, pp. 1–28, 2022.

[32] A. Ilioudi, A. Dabiri, B. J. Wolf, and B. D. Schutter, "Deep learning for object detection and segmentation in videos: Toward an integration with domain knowledge," *IEEE Access*, vol. 10, pp. 34562–34576, 2022.

[33] B. Liu, Y. Guo, J. Jiang, J. Tang, and W. Deng, "Multi-view correlation based black-box adversarial attack for 3D object detection," in *Proc. ACM SIGKDD Conf. Knowl. Disc. Data Min.*, Aug. 2021, pp. 1036–1044.

[34] S. Hu, X. Yuan, W. Ni, X. Wang, and A. Jamalipour, "Visual camouflage and online trajectory planning for unmanned aerial vehicle-based disguised video surveillance: Recent advances and a case study," *IEEE Veh. Technol. Mag.*, vol. 18, no. 3, pp. 48–57, Sep. 2023.

[35] X. Yuan, S. Hu, W. Ni, X. Wang, and A. Jamalipour, "Empowering reconfigurable intelligent surfaces with artificial intelligence to secure air-to-ground Internet-of-Things," *IEEE Int. Things Mag.*, early access.

[36] W. Jiang, Z. He, J. Zhan, W. Pan, and D. Adhikari, "Research progress and challenges on application-driven adversarial examples: A survey," *ACM Trans. Cyber Phys. Syst.*, vol. 5, no. 4, pp. 1–25, 2021.

[37] A. Orth, T. C. Stewart, M. Picard, and M. A. Drouin, "Towards a laser warning system in the visible spectrum using a neuromorphic camera," in *Proc. Int. Conf. Neuromorph. Syst.*, Knoxville, TN, USA, Jul. 2022, pp. 1–4.

[38] X. Yang, W. Liu, S. Zhang, W. Liu, and D. Tao, "Targeted attention attack on deep learning models in road sign recognition," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4980–4990, Mar. 2021.

[39] B. Ma et al., "New cloaking region obfuscation for road network-indistinguishability and location privacy," in *Proc. Int. Symp. Res. Attacks Intrusions Defense*, Limassol, Cyprus, Oct. 2022, pp. 160–170.

[40] I. Fursov et al., "Adversarial attacks on deep models for financial transaction records," in *Proc. ACM SIGKDD Conf. Knowl. Disc. Data Min.*, Aug. 2021, pp. 2868–2878.

[41] S. Khan and M. R. Rabbani, "Chatbot as islamic finance expert (CaIFE): When finance meets artificial intelligence," in *Proc. Int. Symp. Comput. Commun.*, Newcastle upon Tyne, U.K., 2021, pp. 1–5.

[42] G. E. Kaiqiang and T. Chen, "A survey of attack and defense on human–computer interaction security," *Telecommun. Sci.*, vol. 35, no. 10, pp. 100–116, 2019.

[43] H. Park, D. Ahn, K. Hosanagar, and J. Lee, "Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens," in *Proc. ACM CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–15.

[44] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3868–3880, Jun. 2022.

[45] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Trans. Rel.*, vol. 70, no. 1, pp. 389–401, Mar. 2021.

[46] Y. Ye, Y. Chen, and M. Liu, "Multiuser adversarial attack on deep learning for OFDM detection," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2527–2531, Dec. 2022.

[47] S. Zheng et al., "Primary user adversarial attacks on deep learning-based spectrum sensing and the defense method," *China Commun.*, vol. 18, no. 12, pp. 94–107, 2021.

[48] C. Xiao, B. Li, J. Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *Proc. Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 3905–3911.

[49] P. F. de Araujo Filho, G. Kaddoum, M. Naili, E. T. Fapi, and Z. Zhu, "Multi-objective GAN-based adversarial attack technique for modulation classifiers," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1583–1587, Jul. 2022.

[50] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 294–303, Mar. 2021.

[51] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against DNN-based wireless communication systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, New York, NY, USA, 2021, pp. 126–140.

[52] J. Kotak and Y. Elovici, "Adversarial attacks against IoT identification systems," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 7868–7883, May 2023.

[53] P. Huang, X. Zhang, S. Yu, and L. Guo, "IS-WARS: Intelligent and stealthy adversarial attack to Wi-Fi-based human activity recognition systems," *IEEE Trans. Depend. Secure Comput.*, vol. 19, no. 6, pp. 3899–3912, Nov./Dec. 2022.

[54] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Rep.*, San Diego, CA, USA, May 2015, pp. 1–11.

[55] B. Liu, H. Zhang, Y. Wan, F. Zhou, Q. Wu, and D. W. K. Ng, "Robust adversarial attacks on deep learning-based RF fingerprint identification," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1037–1041, Jun. 2023.

[56] Y. Xu, G. Xu, Z. An, M. H. Nielsen, and M. Shen, "Adversarial attacks and active defense on deep learning based identification of GaN power amplifiers under physical perturbation," *AEU Int. J. Electron. Commu.*, vol. 159, Feb. 2023, Art. no. 154478.

[57] F. Xiao, Y. Huang, Y. Zuo, W. Kuang, and W. Wang, "Over-the-air adversarial attacks on deep learning Wi-Fi fingerprinting," *IEEE Internet Things J.*, vol. 10, no. 11, pp. 9823–9835, Jun. 2023.

[58] D. Xu, H. Yang, C. Gu, Z. Chen, Q. Xuan, and X. Yang, "Adversarial examples detection of radio signals based on multifeature fusion," *IEEE Trans. Circuits Syst.*, vol. 68, no. 12, pp. 3607–3611, Dec. 2021.

[59] L. Zhang, S. Lambotharan, G. Zheng, G. Liao, A. Demontis, and F. Roli, "A hybrid training-time and run-time defense against adversarial attacks in modulation classification," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1161–1165, Jun. 2022.

[60] Z. Wang, W. Liu, and H. M. Wang, "GAN against adversarial attacks in radio signal classification," *IEEE Commun. Lett.*, vol. 26, no. 12, pp. 2851–2854, Dec. 2022.

[61] R. Sahay, C. G. Brinton, and D. J. Love, "A deep ensemble-based wireless receiver architecture for mitigating adversarial attacks in automatic modulation classification," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 71–85, Mar. 2022.

[62] K. W. McClintick, J. Harer, B. Flowers, W. C. Headley, and A. M. Wyglinski, "Countering physical eavesdropper evasion with adversarial training," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1820–1833, 2022.

[63] B. R. Manoj, P. M. Santos, M. Sadeghi, and E. G. Larsson, "Toward robust networks against adversarial attacks for radio signal modulation classification," in *Proc. IEEE 23rd Int. Workshop Signal Process. Adv. Wireless Commun.*, 2022, pp. 1–5.

[64] L. Zhang, S. Lambotharan, G. Zheng, G. Liao, B. AsSadhan, and F. Roli, "Attention-based adversarial robust distillation in radio signal classifications for low-power IoT devices," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2646–2657, Feb. 2023.

[65] S. Wang, T. Lv, W. Ni, N. C. Beaulieu, and Y. J. Guo, "Joint resource management for MC-NOMA: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5672–5688, Sep. 2021.

[66] U. Boora, X. Wang, and S. Mao, "Robust massive MIMO localization using neural ODE in adversarial environments," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 4866–4871.

[67] B. R. Manoj, M. Sadeghi, and E. G. Larsson, "Adversarial attacks on deep learning based power allocation in a massive MIMO network," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[68] P. Sun, S. Li, J. Xie, H. Xu, Z. Cheng, and R. Yang, "GPMT: Generating practical malicious traffic based on adversarial attacks with little prior knowledge," *Comput. Security*, vol. 130, Jul. 2023, Art. no. 103257.

[69] A. McCarthy, E. Ghadafi, P. Andriotis, and P. Legg, "Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification," *J. Inf. Security Appl.*, vol. 72, no. 1, 2023, Art. no. 103398.

[70] S. Mohsen, M. Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 86–94.

[71] A. M. Sadeghzadeh, S. Shiravi, and R. Jalili, "Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 2, pp. 1962–1976, Nov. 2021.

[72] E. Nowroozi, Y. Mekdad, M. H. Berenjestanaki, M. Conti, and A. E. Fergougui, "Demystifying the transferability of adversarial attacks in computer networks," *IEEE Trans. Netw. Services Manag.*, vol. 19, no. 3, pp. 3387–3400, Sep. 2022.

[73] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Security Privacy*, Mar. 2016, pp. 372–387.

[74] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Rep.*, Toulon, France, Apr. 2017, pp. 1–17.

[75] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Rep.*, Banff, AB, Canada, Apr. 2014, pp. 1–10.

[76] S. Mohsen, M. Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2574–2582.

[77] R. C. Voicu and Y. Chang, "Cooperative networking using passive discovery for dynamic environments," in *Proc. Int. Wireless Commun. Mobile Comput.*, 2022, pp. 1279–1284.

[78] Z. Yang, J. Cao, Z. Liu, X. Zhang, K. Sun, and Q. Li, "Good learning, bad performance: A novel attack against RL-based congestion control systems," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1069–1082, 2022.

[79] F. Yang, B. Wen, C. Comaniciu, K. P. Subbalakshmi, and R. Chandramouli, "TONet: A fast and efficient method for traffic obfuscation using adversarial machine learning," *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2537–2541, Nov. 2022.

[80] D. Zhan, Y. Duan, Y. Hu, L. Yin, Z. Pan, and S. Guo, "AMGmal: Adaptive mask-guided adversarial attack against malware detection with minimal perturbation," *Comput. Security*, vol. 127, pp. 103103–103115, Apr. 2023.

[81] Q. Cui, Z. Zhu, W. Ni, X. Tao, and P. Zhang, "Edge-intelligence-empowered, unified authentication and trust evaluation for heterogeneous beyond 5G systems," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 78–85, Apr. 2021.

[82] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, and Q. Yu, "A survey of adversarial attack and defense methods for malware classification in cyber security," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 467–496, 1st Quart., 2023.

[83] I. Debicha, R. Bauwens, T. Debatty, J. Dricot, T. Kenaza, and W. Mees, "TAD: Transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems," *Future Gener. Comput. Syst.*, vol. 138, pp. 185–197, Jan. 2023.

[84] H. Jiang, J. Lin, and H. Kang, "FGMD: A robust detector against adversarial attacks in the IoT network," *Future Gener. Comput. Syst.*, vol. 132, pp. 194–210, Jul. 2022.

[85] G. Apruzzese, R. Vladimirov, A. Tastemirova, and P. Laskov, "Wild networks: Exposure of 5G network infrastructures to adversarial examples," *IEEE Trans. Netw. Services Manag.*, vol. 19, no. 4, pp. 5312–5332, Dec. 2022.

[86] H. Lv, M. Wen, R. Lu, and J. Li, "An adversarial attack based on incremental learning techniques for unmanned in 6G scenes," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5254–5264, Jun. 2021.

[87] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Conf. Security Symp.*, 2016, pp. 601–618.

[88] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Security*, New York, NY, USA, 2015, pp. 1322–1333.

[89] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy*, Los Alamitos, CA, USA, May 2017, pp. 3–18.

[90] L. Ashiku and C. Dagli, "Network intrusion detection system using deep learning," *Procedia Comput. Sci.*, vol. 185, no. 1, pp. 239–247, 2021.

[91] X. Yuan, W. Ni, M. Ding, K. Wei, J. Li, and H. V. Poor, "Amplitude-varying perturbation for balancing privacy and utility in federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1884–1897, 2023.

[92] V. Santhanam and L. S. Davis, "A generic improvement to deep residual networks based on gradient flow," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2490–2499, Jul. 2020.

[93] H. Xu, M. Yang, L. Deng, Y. Qian, and C. Wang, "Neutral cross-entropy loss based unsupervised domain adaptation for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4516–4525, 2021.

[94] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.

[95] C. Badue et al., "Self-driving cars: A survey," *Exp. Syst. Appl.*, vol. 165, pp. 113816–113843, Mar. 2021.

[96] D. Förster, T. Bruckschlögl, J. L. Omer, and T. Schipper, "Challenges and directions for automated driving security," in *Proc. Int. Conf. Control Syst. Comput. Sci.*, Ingolstadt, Germany, Dec. 2022, pp. 1–11.

[97] Q. Sun, X. Yao, A. A. Rao, B. Yu, and S. Hu, "Counteracting adversarial attacks in autonomous driving," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 12, pp. 5193–5206, Jan. 2022.

[98] S. Makki, Z. Assaghir, Y. Taher, R. Haque, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019.

[99] L. Chen, Z. Zhang, Q. Liu, L. Yang, Y. Meng, and P. Wang, "A method for online transaction fraud detection based on individual behavior," in *Proc. ACM Turing Award Celebr. Conf. China*, Chengdu, China, May 2019, pp. 1–8.

[100] S. Cao, X. Yang, C. Chen, J. Zhou, X. Li, and Y. Qi, "Titant: Online real-time transaction fraud detection in ant financial," in *Proc. Int. Conf. Very Large Date Bases Endow.*, Los Angeles, CA, USA, Aug. 2019, pp. 2082–2093.

[101] J. Feine, S. Morana, and A. Maedche, *A Chatbot Response Generation System*, Mensch und Comput., Magdebug, Germany, Sep. 2020, pp. 333–341.

[102] W. Seymour, M. Coté, and J. M. Such, "When it's not worth the paper it's written on: A provocation on the certification of skills in the Alexa and Google assistant ecosystems," in *Proc. Conversational User Interfaces*, Glasgow, U.K., Jul. 2022, pp. 1–5.

[103] S. A. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.

[104] F. Z. Qachfar, R. M. Verma, and A. Mukherjee, "Leveraging synthetic data and PU learning for phishing email detection," in *Proc. ACM Conf. Data Appl. Security Privacy*, Baltimore, MD, USA, Apr. 2022, pp. 29–40.

[105] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[106] A. Domyati and Q. Memon, "Robust detection of cardiac disease using machine learning algorithms: Robust detection," in *Proc. Int. Conf. Control Comput. Vis.*, Xiamen, China, Aug. 2022, pp. 52–55.

[107] J. Morris, K. Ergun, B. Khaleghi, M. Imani, B. Aksanli, and T. Simunic, "HyDREA: Utilizing hyperdimensional computing for a more robust and efficient machine learning system," *ACM Trans. Embedded Comput. Syst.*, vol. 21, no. 6, pp. 1–25, 2022.

[108] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Security*, Abu Dhabi, UAE, Apr. 2017, pp. 506–519.

[109] N. Akhtar and A. S. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[110] L. Sun, M. Tan, and Z. Zhou, "A survey of practical adversarial example attacks," *Cybersecurity*, vol. 1, no. 1, pp. 9–17, 2018.

[111] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, pp. 909–939, 2019.

[112] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[113] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, and F. Li, "A survey on adversarial attack in the age of artificial intelligence," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–22, Jun. 2021.

[114] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "Adversarial attack and defense: A survey," *Electronics*, vol. 11, no. 8, pp. 1283–1302, 2022.

[115] Z. Qian, K. Huang, Q. F. Wang, and X. Y. Zhang, "A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies," *Pattern Recognit.*, vol. 131, pp. 108889–108928, Nov. 2022.

[116] A. Amirkhani, M. P. Karimi, and A. Banitalebi-Dehkordi, "A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles," *Vis. Comput.*, to be published.

[117] X. Wei, B. Pu, J. Lu, and B. Wu, "Physically adversarial attacks and defenses in computer vision: A survey," 2022, *arxiv.abs/2211.01671*.

[118] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen, "A survey on physical adversarial attack in computer vision," 2022, *arxiv.abs/2209.14262*.

[119] M. Gallagher, N. Pitropakis, C. Chrysoulas, P. Papadopoulos, A. Mylonas, and S. K. Katsikas, "Investigating machine learning attacks on financial time series models," *Comput. Security*, vol. 123, pp. 102933–102947, Dec. 2022.

[120] M. Goldblum, A. Schwarzschild, A. B. Patel, and T. Goldstein, "Adversarial attacks on machine learning systems for high-frequency trading," in *Proc. ACM Int. Conf. AI Finance*, Nov. 2021, pp. 1–2.

[121] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: High frequency trading in an electronic market," *J. Finance*, vol. 72, no. 3, pp. 967–998, 2017.

[122] N. Carlini and H. Farid, "Evading Deepfake—Image detectors with white- and black-box attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 2804–2813.

[123] Y. Xu, X. Zhong, A. J. Yepes, and J. H. Lau, "Grey-box adversarial attack and defence for sentiment classification," in *Proc. North Amer. Ch. Assoc. Comput. Linguist.*, Jun. 2021, pp. 4078–4087.

[124] W. Ford, "Basic iterative methods," in *Proc. Numer. Linear Algebra Appl.*, 2015, pp. 469–490.

[125] Y. Yu, X. Gao, and C. Z. Xu, "LAFEAT: Piercing through adversarial defenses with latent features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 5735–5745.

[126] Z. Che et al., "Adversarial attack against deep saliency models powered by non-redundant priors," *IEEE Trans. Image Process.*, vol. 30, pp. 1973–1988, 2021.

[127] L. Schwinn et al., "Dynamically sampled nonlocal gradients for stronger adversarial attacks," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[128] G. Yang et al., "Adversarial attack on communication signal modulation recognition," in *Proc. 5th Int. Conf. Inf. Commun. Signal Process.*, 2022, pp. 1–6.

[129] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset DamageNet," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2188–2197, May 2022.

[130] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," in *Proc. IEEE Int. Conf. Big Data Intell. Comput.*, vol. 23, 2019, pp. 828–841.

[131] Z. Wen, Y. Fang, and Z. Liu, "Meta-inductive node classification across graphs," in *Proc. Int. ACM SIGIR Conf. Res. Devices Inf.*, Jul. 2021, pp. 1219–1228.

[132] T. Ugai, "Fuzzy search of knowledge graph with link prediction," in *Proc. Int. Joint Conf. Knowl. Graphs*, Dec. 2021, pp. 121–125.

[133] J. Chen, X. Lin, Z. Shi, and Y. Liu, "Link prediction adversarial attack via iterative gradient attack," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 4, pp. 1081–1094, May 2020.

[134] J. Li, T. Xie, L. Chen, F. Xie, X. He, and Z. Zheng, "Adversarial attack on large scale graph," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 82–95, Jan. 2023.

[135] H. Dai et al., "Adversarial attack on graph structured data," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 1123–1132.

[136] J. Liu, Y. Li, G. Ling, R. Li, and Z. Zheng, "Community detection in location-based social networks: An entropy-based approach," in *Proc. Int. Conf. Comput. Inf. Technol.*, Dec. 2016, pp. 452–459.

[137] J. Wang, M. Luo, F. Suya, J. Li, Z. Yang, and Q. Zheng, "Scalable attack on graph data by injecting vicious nodes," *Data Min. Knowl. Disc.*, vol. 34, no. 5, pp. 1363–1389, 2020.

[138] H. Wang, S. Wang, Z. Jin, Y. Wang, C. Chen, and M. Tistarelli, "Similarity-based gray-box adversarial attack against deep face recognition," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Dec. 2021, pp. 1–8.

[139] H. Wang, G. Li, X. Liu, and L. Lin, "A Hamiltonian Monte Carlo method for probabilistic adversarial attack and learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1725–1737, Apr. 2022.

[140] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[141] Y. Xiang, Y. Xu, Y. Li, W. Ma, Q. Xuan, and Y. Liu, "Side-channel gray-box attack for DNNs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 1, pp. 501–505, Jan. 2021.

[142] Z. Liu, Y. Luo, L. Wu, S. Li, Z. Liu, and S. Z. Li, "Are gradients on graph structure reliable in gray-box attacks?" in *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, Atlanta, GA, USA, 2022, pp. 1360—-1368.

[143] S. Chen, Z. Chen, and D. Wang, "Adaptive adversarial training for meta reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[144] W. Wu, Y. Su, M. R. Lyu, and I. King, "Improving the transferability of adversarial samples with adversarial transformations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9024–9033.

[145] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9185–9193.

[146] Y. Dong, S. Cheng, T. Pang, H. Su, and J. Zhu, "Query-efficient black-box adversarial attacks guided by a transfer-based prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9536–9548, Dec. 2022.

[147] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 15315–15324.

[148] B. Bonnet, T. Furon, and P. Bas, "Generating adversarial images in quantized domains," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 373–385, 2022.

[149] H. Chang et al., "Adversarial attack framework on graph embedding models with limited knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4499–4513, May 2023.

[150] S. Bai, F. Zhang, and P. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognit.*, vol. 110, no. 6, pp. 107637–107644, 2021.

[151] P. Y. Chen, C. C. Tu, P. S. Ting, Y. Y. Lo, D. Koutra, and A. O. Hero, "Identifying influential links for event propagation on Twitter: A network of networks approach," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 1, pp. 139–151, Mar. 2019.

[152] S. Feng, F. Feng, X. Xu, Z. Wang, Y. Hu, and L. Xie, "Digital watermark perturbation for adversarial examples to fool deep neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[153] N. Aafaq, N. Akhtar, W. Liu, M. Shah, and A. Mian, "Language model agnostic gray-box adversarial attack on image captioning," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 626–638, 2023.

[154] Q. Li, Y. Qi, Q. Hu, S. Qi, Y. Lin, and J. S. Dong, "Adversarial adaptive neighborhood with feature importance-aware convex interpolation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2447–2460, 2021.

[155] F. X. Arias, M. Z. Núñez, A. G. Adames, N. T. Flores, and M. V. Lombardo, "Sentiment analysis of public social media as a tool for health-related topics," *IEEE Access*, vol. 10, pp. 74850–74872, 2022.

[156] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 997–1005.

[157] K. Li, W. Ni, X. Yuan, A. Noor, and A. Jamalipour, "Deep-graph-based reinforcement learning for joint cruise control and task offloading for aerial edge Internet of Things (EdgeIoT)," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21676–21686, Nov. 2022.

[158] K. Li et al., "When Internet of Things meets metaverse: Convergence of physical and cyber worlds," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 4148–4173, Mar. 2023.

[159] E. Yang, T. Liu, C. Deng, and D. Tao, "Adversarial examples for hamming space search," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1473–1484, Apr. 2020.

[160] M. Liu, L. Chen, X. Du, L. Jin, and M. Shang, "Activated gradients for deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 2156–2168, Apr. 2023.

[161] S. Jia, Y. Song, C. Ma, and X. Yang, "IoU attack: Towards temporally coherent black-box adversarial attack for visual object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 6709–6718.

[162] Y. Zhang, X. Tian, Y. Li, X. Wang, and D. Tao, "Principal component adversarial example," *IEEE Trans. Image Process.*, vol. 29, pp. 4804–4815, 2020.

[163] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, "AdvDrop: Adversarial attack to DNNs by dropping information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 7486–7495.

[164] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 274–283.

[165] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. Netw. Distrib. Syst. Symp.*, San Diego, CA, USA, Feb. 2018, pp. 1–15.

[166] Z. Cheng et al., "Physical attack on monocular depth estimation with optimal adversarial patches," in *Proc. Eur. Conf. Comput. Vis.*, Oct.2022, pp. 514–532.

[167] M. Shen, H. Yu, L. Zhu, K. Xu, Q. Li, and J. Hu, "Effective and robust physical-world attacks on deep learning face recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4063–4077, 2021.

[168] X. Wei, Y. Guo, and J. Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2711–2725, May 2023.

[169] K. Eykholt et al., "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634.

[170] A. Liu et al., "Perceptual-sensitive GAN for generating adversarial patches," in *Proc. Assoc. Adv. Artif. Intell.*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 1028–1035.

[171] W. Wang et al., "Generating adversarial patches using data-driven multiD-WGAN," in *Proc. IEEE Int. Symp. Circuits Syst.*, Daegu, South Korea, May 2021, pp. 1–5.

[172] J. Wang, A. Liu, X. Bai, and X. Liu, "Universal adversarial patch attack for automatic checkout using perceptual and attentional bias," *IEEE Trans. Image Process.*, vol. 31, pp. 598–611, 2022.

[173] D. Ruta et al., "ALADIN: All layer adaptive instance normalization for fine-grained style similarity," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 11906–11915.

[174] T. Bai, J. Luo, and J. Zhao, "Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9515–9524, May 2022.

[175] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, "Bias-based universal adversarial patch attack for automatic check-out," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 395–410.

[176] C. Yu et al., "Defending against universal adversarial patches by clipping feature norms," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 16414–16422.

[177] T. Kim, H. J. Lee, and Y. M. Ro, "MAP: Multispectral adversarial patch to attack person detection," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, May 2022, pp. 4853–4857.

[178] B. Tarchoun, I. Alouani, A. B. Khalifa, and M. A. Mahjoub, "Adversarial attacks in a multi-view setting: An empirical study of the adversarial patches inter-view transferability," in *Proc. Int. Conf. Cyberworlds*, Caen, France, Sep. 2021, pp. 299–302.

[179] D. Su, H. Zhang, H. Chen, J. Yi, P. Y. Chen, and Y. Gao, "Is robustness the cost of accuracy—A comprehensive study on the robustness of 18 deep image classification models," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 644–661.

[180] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Rep.*, San Diego, CA, USA, May 2015, pp. 1–14.

[181] E. Wallace, M. Stern, and D. Song, "Imitation attacks and defenses for black-box machine translation systems," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Nov. 2020, pp. 5531–5546.

[182] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash, "Efficient adversarial training with transferable adversarial examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, Jun. 2020, pp. 1178–1187.

[183] Y. Wang, Y. Tan, T. Baker, N. Kumar, and Q. Zhang, "Deep fusion: Crafting transferable adversarial examples and improving robustness of industrial artificial intelligence of things," *IEEE Trans. Ind. Informat.*, vol. 19, no. 6, pp. 7480–7488, Jun. 2023.

[184] Y. Zhou, M. Kantarcioglu, and B. Xi, "Exploring the effect of randomness on transferability of adversarial samples against deep neural networks," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 1, pp. 83–99, Jan./Feb. 2023.

[185] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[186] R. Bernhard, P. A. Moëllic, and J. M. Dutertre, "Luring transferable adversarial perturbations for deep neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[187] B. Zi, S. Zhao, X. Ma, and Y. G. Jiang, "Revisiting adversarial robustness distillation: Robust soft labels make student better," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 16423–16432.

[188] X. Yan, Y. Li, T. Dai, Y. Bai, and S. T. Xia, "D2Defend: Dual-domain based defense against adversarial examples," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[189] H. Endo and A. Taguchi, "Color image enhancement by using hue-saturation gradient," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Taipei, Taiwan, Dec. 2019, pp. 1–2.

[190] G. Liu, I. Khalil, and A. Khreishah, "Using single-step adversarial training to defend iterative adversarial examples," in *Proc. ACM Conf. Data Appl. Security Privacy*, Apr. 2021, pp. 17–27.

[191] Z. He, A. S. Rakin, and D. Fan, "Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 588–597.

[192] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–38, 2022.

[193] X. Du and C. M. Pun, "Robust audio patch attacks using physical sample simulation and adversarial patch noise generation," *IEEE Trans. Multimedia*, vol. 24, pp. 4381–4393, 2022.

[194] J. Deng, L. Dong, R. Wang, R. Yang, and D. Yan, "Decision-based attack to speaker recognition system via local low-frequency perturbation," *IEEE Signal Process. Lett.*, vol. 29, pp. 1432–1436, 2022.

[195] Y. Ding, H. Lin, L. Liu, Z. Ling, and Y. Hu, "Robustness of speech spoofing detectors against adversarial post-processing of voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3415–3426, 2021.

[196] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, "Towards a robust deep neural network against adversarial texts: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3159–3179, Mar. 2023.

[197] S. Liu, N. Lu, C. Chen, and K. Tang, "Efficient combinatorial optimization for word-level adversarial textual attack," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 98–111, 2022.

[198] B. R. Manoj, M. Sadeghi, and E. G. Larsson, "Downlink power allocation in massive MIMO via deep learning: Adversarial attacks and training," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 707–719, Jun. 2022.

[199] P. M. Santos, B. R. Manoj, M. Sadeghi, and E. G. Larsson, "Universal adversarial attacks on neural networks for power allocation in a massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 67–71, Jan. 2022.

[200] Y. Guo, Q. Li, W. Zuo, and H. Chen, "An intermediate-level attack framework on the basis of linear regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2726–2735, Mar. 2023.

[201] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2019, pp. 1–11.

[202] O. Suciu, S. E. Coull, and J. Johns, "Exploring adversarial examples in malware detection," in *Proc. IEEE Security Privacy Workshops*, 2019, pp. 8–14.

[203] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, and F. Roli, "Adversarial Examples: A survey and experimental evaluation of practical attacks on machine learning for windows malware detection," *ACM Trans. Privacy Security*, vol. 24, no. 4, pp. 1–31, Sep. 2021.

[204] J. Kotak and Y. Elovici, "IoT device identification using deep learning," in *Proc. 13th Int. Conf. Comput. Intell. Security Inf. Syst.*, Burgos, Spain, Sep. 2020, pp. 76–86.

[205] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 213–216, Feb. 2019.

[206] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1102–1113, 2020.

[207] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 76–81, May 2019.

[208] Z. Luo, S. Zhao, Z. Lu, J. Xu, and Y. E. Sagduyu, "When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1892–1908, May 2022.

[209] Z. Bao, Y. Lin, S. Zhang, Z. Li, and S. Mao, "Threat of adversarial attacks on DL-based IoT device identification," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9012–9024, Jun. 2022.

[210] D. Han et al., "DeepAID: Interpreting and improving deep learning-based anomaly detection in security applications," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, New York, NY, USA, 2021, pp. 3197–3217.

[211] G. Xie, Q. Li, and Y. Jiang, "Self-attentive deep learning method for online traffic classification and its interpretability," *Comput. Netw.*, vol. 196, pp. 108267–108278, Sep. 2021.

[212] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *Proc. 54th Annu. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 2020, pp. 1–6.

[213] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. 5th Int. Conf. Learn. Rep.*, Toulon, France, Apr. 2017, pp. 1–9.

[214] F. Feng, X. He, J. Tang, and T. S. Chua, "Graph adversarial training: Dynamically regularizing based on graph structure," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2493–2504, Jun. 2021.

[215] S. Amini and S. Ghaemmaghami, "Towards improving robustness of deep neural networks to adversarial perturbations," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1889–1903, Jul. 2020.

[216] Y. Wang, T. Li, S. Li, X. Yuan, and W. Ni, "New adversarial image detection based on sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 19, 2023, doi: 10.1109/TNNLS.2023.3274538.

[217] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Rep.*, Toulon, France, Apr. 2017, pp. 1–12.

[218] H. Li, G. Li, and Y. Yu, "ROSA: Robust salient object detection against adversarial attacks," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4835–4847, Nov. 2020.

[219] K. Han, Y. Li, and B. Xia, "A cascade model-aware generative adversarial example detection method," *Tsinghua Sci. Technol.*, vol. 26, no. 6, pp. 800–812, 2021.

[220] C. P. Ngo, A. A. Winarto, C. K. L. Kou, S. Park, F. Akram, and H. K. Lee, "Fence GAN: Towards better anomaly detection," in *Proc. Int. Conf. Tools Artif. Intell.*, Portland, OR, USA, Nov. 2019, pp. 141–148.

[221] C. Xie, K. Yang, A. Wang, C. Chen, and W. Li, "A MURA detection method based on an improved generative adversarial network," *IEEE Access*, vol. 9, pp. 68826–68836, 2021.

[222] X. Zhang and Y. Wang, "ADNet: A neural network model for adversarial example detection based on steganalysis and attention mechanism," in *Proc. Int. Conf. Comput. Vis. Image Deep Learn.*, San Diego, CA, USA, Oct./Nov. 2021, pp. 55–60.

[223] S. Freitas, S. T. Chen, Z. J. Wang, and D. H. Chau, "UnMask: Adversarial detection and defense through robust feature alignment," in *Proc. IEEE Int. Conf. Big Data*, Atlanta, GA, USA, Dec. 2020, pp. 1081–1088.

[224] Y. Wang, L. Xie, X. Liu, J. Yin, and T. Zheng, "Model-agnostic adversarial example detection through logit distribution learning," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 3617–3621.

[225] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Detection of adversarial attacks and characterization of adversarial subspace," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, May 2020, pp. 3097–3101.

[226] H. Liu, F. Wang, and J. Du, "Research on adversarial attack technology for object detection in physical world based on vision," in *Proc. Asia Conf. Algorithms Comput. Mach. Learn.*, Hangzhou, China, Mar. 2022, pp. 638–648.

[227] S. Wang, "Joint learning of discriminative metric space from multicontext visual scene for unsupervised salient object detection," *IEEE Access*, vol. 10, pp. 126089–126099, 2022.

[228] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," in *Proc. Int. Conf. Learn. Rep.*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–20.

[229] Z. Liu et al., "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 860–868.

[230] X. Ling et al., "DEEPSEC: A uniform platform for security analysis of deep learning model," in *Proc. IEEE Symp. Security Privacy*, San Francisco, CA, USA, May 2019, pp. 673–690.

[231] M. Kang, T. Q. Tran, S. J. Cho, and D. Kim, "CAP-GAN: Towards adversarial robustness with cycle-consistent attentional purification," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[232] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of JPG compression on adversarial images," 2016, *arxiv.abs/1608.00853*.

[233] D. Zhou et al., "Removing adversarial noise in class activation feature space," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 7858–7867.

[234] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "APE-GAN: Adversarial perturbation elimination with GAN," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., May 2019, pp. 3842–3846.

[235] W. Wei and L. Liu, "Robust deep learning ensemble against deception," *IEEE Trans. Depend. Security Comput.*, vol. 18, no. 4, pp. 1513–1527, Jul./Aug. 2021.

[236] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. V. Nori, and A. Criminisi, "Measuring neural net robustness with constraints," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 2613–2621.

[237] C. Etmann, S. Lunz, P. Maass, and C. Schönlieb, "On the connection between adversarial robustness and saliency map interpretability," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, vol. 97, Jun. 2019, pp. 1823–1832.

[238] M. Jordan, J. Lewis, and A. G. Dimakis, "Provable certificates for adversarial examples: Fitting a ball in the union of Polytopes," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 14059–14069.

[239] Y. Zhu, X. Wei, and Y. Zhu, "Efficient adversarial defense without adversarial training: A batch normalization approach," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[240] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Trans. Image Process.*, vol. 29, pp. 1711–1724, 2020.

[241] T. Blau, R. Ganz, B. Kawar, A. M. Bronstein, and M. Elad, "Threat model-agnostic adversarial defense using diffusion models," 2022, *arxiv.abs/2207.08089*.

[242] Y. Qin and C. Yue, "Key-based input transformation defense against adversarial examples," in *Proc. Int. Perform. Comput. Commun. Conf.*, Austin, TX, USA, Oct. 2021, pp. 1–10.

[243] A. D. Marchi and A. Themelis, "Proximal gradient algorithms under local Lipschitz gradient continuity," *J. Optim. Theory Appl.*, vol. 194, no. 3, pp. 771–794, 2022.

[244] R. Anirudh, J. J. Thiagarajan, B. Kailkhura, and P. T. Bremer, "MimicGAN: Robust projection onto image manifolds with corruption mimicking," *Int. J. Comput. Vis.*, vol. 128, no. 10, pp. 2459–2477, 2020.

[245] J. Xiao, S. Zhang, Y. Yao, Z. Wang, Y. Zhang, and Y. F. Wang, "Generative adversarial network with hybrid attention and compromised normalization for multi-scene image conversion," *Neural Comput. Appl.*, vol. 34, no. 9, pp. 7209–7225, 2022.

[246] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arxiv.abs/1706.03825*.

[247] Z. Yan, Y. Guo, and C. Zhang, "Adversarial margin maximization networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1129–1139, Apr. 2021.

[248] C. Yang, L. Ding, Y. Chen, and H. Li, "Defending against GAN-based deepFake attacks via transformation-aware adversarial faces," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[249] H. Yu, A. Liu, G. Li, J. Yang, and C. Zhang, "Progressive diversified augmentation for general robustness of DNNs: A unified approach," *IEEE Trans. Image Process.*, vol. 30, pp. 8955–8967, 2021.

[250] S. Kundu, M. Pedram, and P. A. Beerel, "HIRE-SNN: Harnessing the adversarial robustness of energy-efficient deep spiking neural networks via training with crafted input noise," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 5189–5198.

[251] Z. Xu, J. Wang, and J. Pu, "Defense against adversarial attacks with an induced class," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[252] R. Yao, C. Huang, Z. Hu, and K. Pei, "Adaptive retraining for neural network robustness in classification," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[253] C. Gong, T. Ren, M. Ye, and Q. Liu, "MaxUp: Lightweight adversarial training with data augmentation improves neural network training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2474–2483.

[254] Q. Liu, C. Long, J. Zhang, M. Xu, and P. Lv, "TriATNE: Tripartite adversarial training for network embeddings," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9634–9645, Sep. 2022.

[255] O. Poursaeed, T. Jiang, H. Yang, S. J. Belongie, and S. N. Lim, "Robustness and generalization via generative adversarial training," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 15691–15700.

[256] M. Hong, J. Choi, and G. Kim, "StyleMix: Separating content and style for enhanced data augmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 14862–14870.

[257] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct./Nov. 2019, pp. 6022–6031.

[258] H. Shen, S. Chen, R. Wang, and X. Wang, "Adversarial learning with cost-sensitive classes," *IEEE Trans. Cybern.*, vol. 53, no. 8, pp. 4855–4866, Aug. 2023.

[259] G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian, "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 448–457.

[260] M. Sun et al., "Can shape structure features improve model robustness under diverse adversarial settings?" in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 7506–7515.

[261] A. Liu, X. Liu, H. Yu, C. Zhang, Q. Liu, and D. Tao, "Training robust deep neural networks via adversarial noise propagation," *IEEE Trans. Image Process.*, vol. 30, pp. 5769–5781, 2021.

[262] Y. Dong, H. Wang, and Y. Yao, "A robust adversarial network-based end-to-end communications system with strong generalization ability against adversarial attacks," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 4086–4091.

[263] C. Zhang et al., "Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity," *IEEE Trans. Image Process.*, vol. 30, pp. 1291–1304, 2021.

[264] D. M. Schwartz and G. Ditzler, "Bolstering adversarial robustness with latent disparity regularization," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[265] Y. Jiang, X. Ma, S. M. Erfani, and J. Bailey, "Dual head adversarial training," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[266] Y. Li et al., "Towards robustness of deep neural networks via Regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 7476–7485.

[267] G. Carbone, G. Sanguinetti, and L. Bortolussi, "Random projections for improved adversarial robustness," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–7.

[268] C. R. Genovese, M. Perone-Pacifico, and V. L. Wasserman, "Manifold estimation and singular deconvolution under Hausdorff loss," *Ann. Stat.*, vol. 40, no. 2, pp. 941–963, 2012.

[269] S. Dasgupta and Y. Freund, "Random projection trees and low dimensional manifolds," in *Proc. ACM Symp. Theory Comput.*, May 2008, pp. 537–546.

[270] H. Liu, H. Wu, W. Xie, F. Liu, and L. Shen, "Group-wise inhibition based feature regularization for robust classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 468–476.

[271] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 8671–8685, 2021.

[272] Q. Li, L. Shen, S. Guo, and Z. Lai, "WaveCNet: Wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 7074–7089, 2021.

[273] J. Mok, B. Na, H. Choe, and S. Yoon, "AdvRush: Searching for adversarially robust neural architectures," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 12302–12312.

[274] T. Yeo, O. Kar, and A. Zamir, "Robustness via cross-domain ensembles," in *Proc. IEEE Int. Conf. Comput. Vis.*, Los Alamitos, CA, USA, 2021, pp. 12169–12179.

[275] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 2266–2276.

[276] M. Serrurier, F. Mamalet, A. G. Sanz, T. Boissin, J. M. Loubes, and E. D. Barrio, "Achieving robustness in classification using optimal transport with hinge regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 505–514.

[277] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *Proc. Int. Conf. Learn. Rep.*, San Diego, CA, USA, May 2015, pp. 1–9.

[278] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *Proc. Assoc. Adv. Artif. Intell.*, Feb./Mar. 2022, pp. 1306–1313.

[279] H. Wang, Y. Deng, S. Yoo, H. Ling, and Y. Lin, "AGKD-BML: Defense against adversarial attack by attention guided knowledge distillation and bi-directional metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7638–7647.

[280] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 478–489.

[281] J. Wang and H. Zhang, "Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct./Nov. 2019, pp. 6628–6637.

[282] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 1829–1839.

[283] M. Kuzlu, F. O. Catak, U. Cali, E. Catak, and O. Guler, "Adversarial security mitigations of mmWave beamforming prediction models using defensive distillation and adversarial retraining," *Int. J. Inf. Security*, vol. 22, no. 2, pp. 319–332, Nov. 2022.

[284] B. Wu, S. Wang, X. Yuan, C. Wang, C. Rudolph, and X. Yang, "Defeating misclassification attacks against transfer learning," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 2, pp. 886–901, Mar./Apr. 2023.

[285] N. Alon, A. Gonen, E. Hazan, and S. Moran, "Boosting simple learners," in *Proc. ACM Symp. Theory Comput.*, Jun. 2021, pp. 481–489.

[286] R. Kumar and G. Subbiah, "Explainable machine learning for malware detection using ensemble bagging algorithms," in *Proc. Int. Conf. Contemp. Comput.*, Aug. 2022, pp. 453–460.

[287] A. Mustafa, S. H. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Deeply supervised discriminative learning for adversarial defense," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3154–3166, Sep. 2021.

[288] A. Shafahi, A. Ghiasi, M. Najibi, F. Huang, J. P. Dickerson, and T. Goldstein, "Batch-wise Logit-similarity: Generalizing Logit-squeezing and label-smoothing," in *Proc. Brit. Mach. Vis. Conf.*, Cardiff, U.K., Sep. 2019, pp. 72–83.

[289] S. Kanai, M. Yamada, S. Yamaguchi, H. Takahashi, and Y. Ida, "Constraining logits by bounded function for adversarial robustness," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, Jul. 2021, pp. 1–8.

[290] C. Liu, M. Salzmann, and S. Süsstrunk, "Training provably robust models by polyhedral envelope regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 3146–3160, Jun. 2023.

[291] M. Balunovic and M. T. Vechev, "Adversarial training and provable defenses: Bridging the gap," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2020, pp. 1–19.

[292] H. Lee, H. Bae, and S. Yoon, "Gradient masking of label smoothing in adversarial robustness," *IEEE Access*, vol. 9, pp. 6453–6464, 2021.

[293] A. Howard et al., "Searching for mobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct./Nov. 2019, pp. 1314–1324.

[294] P. Yun, Y. Liu, and M. Liu, "In defense of knowledge distillation for task incremental learning and its application in 3D object detection," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2012–2019, Apr. 2021.

[295] X. Liu, H. Jiang, X. Su, and J. Feng, "Adversarial examples generation algorithm based on decision boundary search," *J. Electron. Sci. Technol.*, vol. 51, pp. 721–727, May 2022.

[296] G. Yu, X. Wang, P. Yu, C. Sun, W. Ni, and R. P. Liu, "Dataset obfuscation: Its applications to and impacts on edge machine learning," *ACM Trans. Intell. Syst. Technol.*, to be published.

[297] S. Hu, X. Chen, W. Ni, E. Hossain, and X. Wang, "Distributed machine learning for wireless communication networks: Techniques, architectures, and applications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1458–1493, 3rd Quart., 2021.

[298] J. Zheng, H. Tian, W. Ni, W. Ni, and P. Zhang, "Balancing accuracy and integrity for reconfigurable intelligent surface-aided over-the-air federated learning," *IEEE Trans. Wireless Commu.*, vol. 21, no. 12, pp. 10964–10980, Dec. 2022.

[299] R. Saleem, W. Ni, M. Ikram, and A. Jamalipour, "Deep-reinforcement-learning-driven secrecy design for intelligent-reflecting-surface-based 6G-IoT networks," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8812–8824, May 2023.

[300] S. Zhang, H. Yin, T. Chen, Q. V. N. Hung, Z. Huang, and L. Cui, "GCN-based user representation learning for unifying robust recommendation and Fraudster detection," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, New York, NY, USA, 2020, pp. 689–698.

[301] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. 7th Int. Conf. Learn. Rep.*, New Orleans, LA, USA, May 2019, pp. 1–9.

[302] E. Wegert and M. A. Efendiev, "Nonlinear Riemann—Hilbert problems with Lipschitz continuous boundary condition," *Rensselaer Soc. Eng.*, vol. 130, no. 04, pp. 793–800, 2000.

[303] X. Chen, Z. S. Wu, and M. Hong, "Understanding gradient clipping in private SGD: A geometric perspective," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 13773–13782.

[304] E. Lobacheva, M. Kodryan, N. Chirkova, A. Malinin, and D. P. Vetrov, "On the periodic Behavior of neural network training with batch normalization and weight decay," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 21545–21556.

[305] Y. Du, W. Cao, J. She, M. Wu, M. Fang, and S. Kawata, "Disturbance rejection and control system design using improved equivalent input disturbance approach," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 3013–3023, Apr. 2020.

[306] P. Dhilleswararao, S. Boppu, M. S. Manikandan, and L. R. Cenkeramaddi, "Efficient hardware architectures for accelerating deep neural networks: Survey," *IEEE Access*, vol. 10, pp. 131788–131828, 2022.

[307] T. Zhang, K. Yang, and J. Wei, "Survey on detecting and defending adversarial examples for image data," *J. Comput. Res. Dev.*, vol. 59, no. 6, pp. 1315–1328, 2022.

[308] M. Usama, M. Asim, S. Latif, J. Qadir, and A. I. A. Fuqaha, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *Proc. Wireless Commun. Mobile Comput.*, Tangier, Morocco, Jun. 2019, pp. 78–83.

Yulong Wang (Member, IEEE) received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications, China, in 2010, where he is currently an Associate Professor and a Ph.D. Supervisor with the School of Computer Science (National Pilot Software Engineering School) and the State Key Laboratory of Networking and Switching Technology. He was a Visiting Scientist with CSIRO, Australia, from 2019 to 2020. His research interests include deep learning, Internet of Things, large language model, and network security.

Tong Sun received the B.E. degree in Internet of Things engineering from the Nanjing University of Posts and Telecommunications in 2017. She is currently pursuing the master's degree in computer science and technology with the Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include deep learning, adversarial learning, and network security.

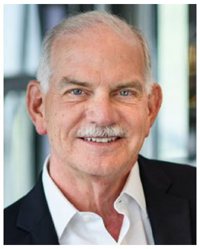
Shenghong Li (Member, IEEE) received the B.S. degree in communication engineering from Nanjing University, Nanjing, Jiangsu, China, in 2008, and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a Senior Research Scientist with Data61, CSIRO. His research interests include computer vision, deep learning, wireless tracking, data fusion, and wireless communication.

Xin Yuan (Member, IEEE) received the B.E. degree from the Taiyuan University of Technology, Shanxi, China, in 2013, and the dual Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, and the University of Technology Sydney, Sydney, NSW, Australia, in 2019 and 2020, respectively. She is currently a Research Scientist with CSIRO, Sydney. Her research interests include machine learning and optimization, and their applications to Internet of Things and intelligent systems.

Wei Ni (Senior Member, IEEE) received the B.E. and Ph.D. degrees in communication science and engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively.

He is currently a Principal Research Scientist with CSIRO, Sydney, Australia, a Conjoint Professor with the University of New South Wales, an Adjunct Professor with the University of Technology Sydney, and an Honorary Professor with Macquarie University. He was a Postdoctoral Research Fellow with Shanghai Jiaotong University from 2005 to 2008; a Deputy Project Manager with the Bell Labs, Alcatel/Alcatel-Lucent from 2005 to 2008; and a Senior Researcher with Devices R&D, Nokia from 2008 to 2009. He has authored seven book chapters, more than 280 journal papers, 100 conference papers, 25 patents, and ten standard proposals accepted by IEEE. His research interests include machine learning, online learning, stochastic optimization, and their applications to system efficiency and integrity. He has been the Chair of IEEE Vehicular Technology Society (VTS) New South Wales (NSW) Chapter since 2020, an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018, and an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He served first as the Secretary and then the Vice-Chair of IEEE NSW VTS Chapter from 2015 to 2019, the Track Chair for VTC-Spring 2017, the Track Co-Chair for IEEE VTC-Spring 2016, the Publication Chair for BodyNet 2015, and the Student Travel Grant Chair for WPMC 2014.

Ekram Hossain (Fellow, IEEE) is a Professor and the Associate Head (Graduate Studies) with the Department of Electrical and Computer Engineering, University of Manitoba, Canada. His current research interests include design, analysis, and optimization of wireless networks with emphasis on beyond 5G cellular networks. He received the 2017 IEEE ComSoc Technical Committee on Green Communications and Computing Distinguished Technical Achievement Recognition Award "for outstanding technical leadership and achievement in green wireless communications and networking." He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2017–2022. He served as the Editor-in-Chief of the IEEE PRESS from 2018 to 2021 and the Editor-in-Chief of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2012 to 2016. He currently serves as the Director of Online Content of IEEE ComSoc in 2022–2023. He was an Elected Member of the Board of Governors of the ComSoc from 2018 to 2020 and served as the Director of Magazines for IEEE ComSoc from 2020 to 2021. He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks." He is a member (Class of 2016) of the College of the Royal Society of Canada, and a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He received the IEEE Alexander Graham Bell Medal in 2017. He is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies.