Deep Learning Systems in the Physical World: A Survey on Adversarial Robustness and Real-World Applications

www.surveyx.cn

Abstract

This survey paper explores the integration of deep learning systems within the physical world, emphasizing the critical need to enhance adversarial robustness for secure and reliable real-world applications. Deep learning models, while transformative in domains such as autonomous driving and surveillance, are vulnerable to adversarial attacks—malicious inputs designed to mislead models. These vulnerabilities are particularly concerning in safety-critical applications, where adversarial examples can induce erroneous outputs, compromising system integrity. The survey highlights significant advancements in defense strategies, such as self-supervised adversarial training and innovative frameworks like Jujutsu and LanCe, which improve model resilience against evolving adversarial threats. The potential of large language models in analyzing physical signals is also explored, underscoring the expanding role of AI in cyber-physical systems. The paper emphasizes the importance of developing adaptive and resilient defense mechanisms that account for environmental variability and cyber-physical interactions. By reviewing methodologies for robustness evaluation, including perceptual metrics and standardized benchmarks, the survey provides a comprehensive understanding of the challenges and future directions in enhancing adversarial robustness. The findings underscore the necessity for continued research and innovation to ensure the secure deployment of deep learning systems across diverse real-world scenarios.

1 Introduction

1.1 Deep Learning Systems in the Physical World

Deep learning systems are pivotal in enhancing various applications in the physical world, particularly within Cyber-Physical Systems (CPS), where they improve security and operational efficiency [1]. These systems find applications in autonomous vehicles, surveillance, and face recognition, processing complex environmental data for informed decision-making. For instance, autonomous vehicles utilize deep neural networks (DNNs) to interpret sensor data essential for lane detection and trajectory prediction [2]. However, their deployment faces significant challenges, notably vulnerability to adversarial attacks and environmental variability [3].

The susceptibility of deep learning systems to adversarial examples poses a critical concern, as minimal input perturbations can mislead classifiers [3]. In face recognition, models are particularly vulnerable to adversarial images in physical contexts, where traditional defenses like print or replay attacks are insufficient [4]. Similarly, DNN applications in aerial detection are compromised by adversarial attacks, presenting real-world risks [5].

Current adversarial patch methods require extensive computational resources and detailed model knowledge, limiting their practical applicability in real-world scenarios [6]. Monocular Depth Estimation (MDE), critical for autonomous driving, is also threatened by adversarial attacks, highlighting

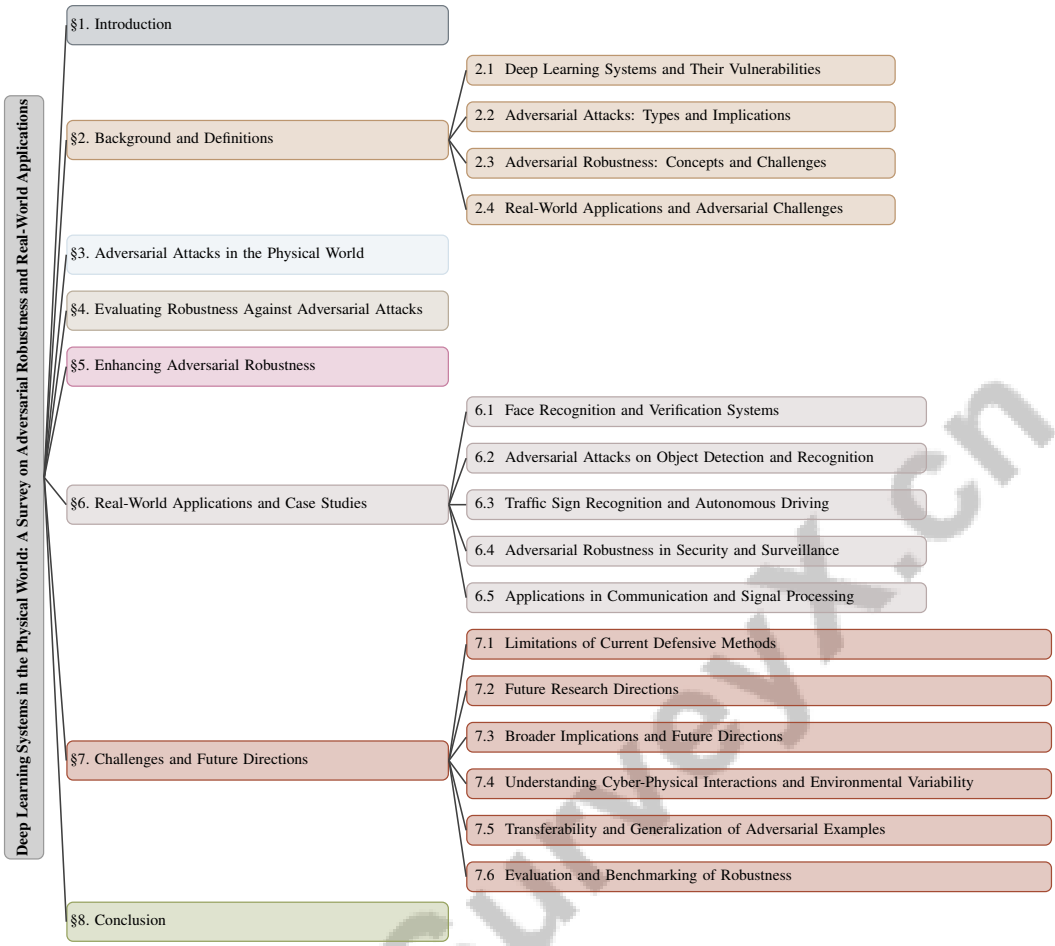


Figure 1: chapter structure

the urgent need for robust defenses [7]. These vulnerabilities underscore the necessity for ongoing research to enhance the robustness of deep learning systems, ensuring their secure and reliable operation in physical environments.

1.2 Importance of Adversarial Robustness

Adversarial robustness is essential for the reliable operation of deep learning systems, especially in safety-critical applications. The prevalence of adversarial attacks in domains such as autonomous vehicles and healthcare underscores the need for robust defenses to ensure system reliability [8]. These attacks can mislead models while appearing visually natural, compromising the performance and safety of deep learning systems [9]. The vulnerability of face authentication systems to adversarial images in physical environments further illustrates the critical need for adversarial robustness, as these attacks can bypass existing Presentation Attack Detectors (PADs) [4].

Defending against localized adversarial patches remains a significant challenge, emphasizing the need for improved adversarial robustness [10]. In autonomous vehicles, maintaining accurate predictions under adversarial conditions is crucial to prevent misclassifications that could jeopardize safety [11]. Aerial detection systems are similarly threatened by adversarial attacks that manipulate detection outcomes without obstructing targets, further highlighting the importance of robust defenses [5].

Benchmarking the robustness of machine learning classifiers against adversarial examples in physical scenarios is vital for comparing models and methods [3]. This benchmarking aids in identifying vulnerabilities and developing more resilient systems. Additionally, the complexity of training deep learning models on diverse datasets and the lack of interpretability in black-box models present further challenges to achieving adversarial robustness [1].

www.surveyx.cn

AI-generated, for reference only.

Innovative approaches, such as using view synthesis to simulate physical attacks during training, present promising avenues for enhancing adversarial robustness without relying on ground-truth data [7]. Addressing these multifaceted challenges is crucial for the secure and reliable deployment of deep learning systems across various real-world applications.

1.3 Relevance of the Study

Research on adversarial robustness is critical in real-world applications, where deep learning systems are increasingly utilized. Assessing the generalizability of testing results from simulated environments to actual scenarios is essential for ensuring system reliability and security. Empirical studies, such as those by Stocco et al., emphasize the importance of bridging this gap, underscoring the relevance of adversarial robustness research in practical applications [12].

The introduction of Assistive Signals, as proposed by Pestana et al., offers a novel method to enhance prediction confidence, vital for the dependable operation of deep learning models in real-world contexts [13]. This study highlights the need for mechanisms that bolster system resilience against adversarial threats.

Moreover, Chakraborty et al.'s comprehensive survey provides insights into various adversarial attacks and their countermeasures, addressing a critical knowledge gap in adversarial learning [14]. Understanding these threats and developing effective defenses is essential for safeguarding systems against potential adversarial exploits.

The introduction of TnTs, described by Doan et al., presents a new category of adversarial examples that exploit vulnerabilities in deep neural networks through naturalistic patches [15]. Such advancements necessitate ongoing research to develop robust defenses capable of countering these sophisticated threats.

Lastly, the concept of Penetrative AI, which emphasizes integrating large language models (LLMs) with IoT systems as articulated by Xu et al., highlights the expanding role of AI in real-world applications [16]. This integration further accentuates the need for adversarial robustness to ensure secure and effective system functionality across diverse environments.

1.4 Structure of the Survey

This survey is meticulously structured to comprehensively explore deep learning systems in the physical world, focusing on adversarial robustness and real-world applications. The paper is organized into several key sections, each addressing critical aspects of the topic.

The **Introduction** section establishes the significance of deep learning systems in the physical world and the essential role of adversarial robustness, emphasizing the importance of ensuring reliable and secure performance in real-world applications.

The **Background and Definitions** section delves into foundational concepts, providing definitions and explanations of deep learning systems, adversarial attacks, adversarial robustness, and real-world applications, establishing a clear understanding of their interactions within the physical context.

The **Adversarial Attacks in the Physical World** section examines various types of adversarial attacks targeting deep learning systems, discussing the challenges and implications these attacks pose to system performance and security.

Following this, the **Evaluating Robustness Against Adversarial Attacks** section reviews methodologies and frameworks for assessing the robustness of deep learning systems, highlighting key metrics and benchmarks used in evaluations.

In the **Enhancing Adversarial Robustness** section, strategies and techniques employed to bolster adversarial robustness are examined, encompassing both theoretical approaches and practical implementations aimed at fortifying systems against adversarial threats.

The **Real-World Applications and Case Studies** section presents case studies and examples of deep learning systems deployed in various real-world applications, analyzing how adversarial robustness is tested and ensured, providing practical insights into encountered challenges and solutions.

AI-generated, for reference only.

Finally, the **Challenges and Future Directions** section identifies ongoing challenges in achieving adversarial robustness and proposes potential future research directions, exploring broader implications and suggesting innovations to advance the field.

The survey concludes with a **Conclusion** section that summarizes key findings and emphasizes the importance of continued research and development in enhancing the adversarial robustness of deep learning systems for secure and reliable real-world applications.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Deep Learning Systems and Their Vulnerabilities

Deep learning systems, particularly those utilizing deep neural networks (DNNs), are essential in applications like traffic sign recognition, autonomous driving, and object detection. These systems, while proficient in processing complex data, are susceptible to adversarial attacks that exploit their vulnerabilities, leading to erroneous outputs [17]. In traffic sign recognition, adversarial examples can mislead classifiers, posing significant safety risks for autonomous vehicles [18]. Similarly, DNN-based object detection systems face threats from adversarial patches that can obscure targets, undermining reliability [19].

The integrity of these systems is further threatened by backdoor attacks, where adversaries manipulate training data to introduce backdoor instances that misclassify inputs, as seen in robotic manipulation applications [20]. Physical backdoor attacks often utilize poisoned inputs with incorrect labels, increasing detectability [9]. This vulnerability extends to less-explored thermal infrared object detection systems [21]. In safety-critical domains like autonomous vehicles and medical imaging, the susceptibility of DNNs to adversarial attacks is concerning, exacerbated by the challenge of providing formal guarantees about DNN behavior [3]. Perception systems face security challenges from simultaneous physical-world attacks across all sensor sources, such as cameras and LiDAR [2].

The complexity of adversarial threats necessitates effective, stealthy, and robust defenses to ensure the secure operation of deep learning systems in real-world contexts. Existing methods, such as sticker-based attacks, often lack flexibility and can be conspicuous [6]. The primary challenge lies in the covert nature of physical perturbations, with many methods either leaving visible marks or failing in well-lit environments [22]. Deploying these systems involves selecting appropriate algorithms, managing data biases, addressing model interpretability, and ensuring robustness against concept drift [1]. The emergence of realistic, physical-world-resilient adversarial examples complicates the threat landscape, effectively targeting common scenarios in autonomous driving [23]. Addressing these vulnerabilities is essential for safe deployment across various domains.

2.2 Adversarial Attacks: Types and Implications

Adversarial attacks significantly threaten the integrity and performance of deep learning systems, particularly in physical-world applications. These attacks exploit model vulnerabilities by introducing subtle perturbations to input data, often imperceptible to humans, misleading models into incorrect outputs [4]. The scope of these attacks ranges from digital manipulations to complex physical scenarios, where environmental factors can significantly affect execution.

A notable category involves adversarial patches, crafted perturbations designed to deceive image recognition systems while remaining inconspicuous. However, real-world transformations like changes in pose, lighting, and fabric deformation often diminish their effectiveness, complicating applicability in dynamic environments [24]. Creating effective camouflage patterns for complex, non-rigid, or non-planar objects presents further challenges [4]. Innovative strategies include leveraging natural phenomena like shadows as perturbations, achieving high success rates in real-world scenarios such as traffic sign recognition [9, 25, 26, 27]. This highlights the potential for utilizing everyday environmental features in crafting subtle yet effective adversarial attacks, necessitating robust defenses against such vulnerabilities.

Backdoor attacks embed hidden triggers within training data, which can be activated in real-world scenarios to manipulate model behavior. The difficulty of adapting traditional lane detection methods, primarily focused on classification tasks, to non-classification challenges like lane detection

AI-generated, for reference only.

underscores the intricate challenges posed by adversarial threats in autonomous driving [28, 22]. The current lack of effective strategies for backdoor attacks on systems such as Synthetic Aperture Radar (SAR) reveals a gap in methodologies that predominantly focus on digital rather than physical environments.

The implications of adversarial attacks are profound, jeopardizing model performance and raising concerns about safety and reliability. These attacks exploit inherent vulnerabilities, leading to misclassifications and undermining trust in applications dependent on these technologies. As adversaries increasingly employ sophisticated techniques—such as model extraction, inversion, and poisoning attacks—developing robust countermeasures becomes imperative to safeguard the integrity of deep learning systems across various domains [29, 14, 30]. The limitations of existing physical attack methods, often too visible or ineffective across different angles and distances, highlight the need for more sophisticated techniques adaptable to real-world conditions.

2.3 Adversarial Robustness: Concepts and Challenges

Adversarial robustness in deep learning systems is the ability of models to maintain accuracy and reliability when confronted with adversarial examples designed to mislead them [31]. This robustness is crucial in safety-sensitive applications like autonomous driving, where adversarial attacks can significantly compromise system performance through physical adversarial patches [2]. Achieving adversarial robustness is challenging due to the multifaceted nature of adversarial threats and the limitations of current defense strategies.

A primary challenge is the susceptibility of deep learning models to universal adversarial attacks, which exploit uncertainties in model decision-making processes [32]. These attacks can be executed physically, utilizing adversarial patches or textures applied to objects designed to deceive models across various viewing conditions [17]. Ensuring adversarial robustness is further complicated by the need for perturbations to remain effective across multiple viewpoints while preserving a natural appearance, as evidenced by the challenges in developing physically realizable adversarial attacks [23].

Current defenses against adversarial patches often rely on heuristic methods that lack robustness against adaptive attackers, highlighting the necessity for more effective solutions [10]. Innovative approaches, such as the Detector Collapse (DC) method, which introduces backdoors to disable object detectors by exploiting both regression and classification branches, exemplify the evolving nature of adversarial threats [19]. The introduction of black-box attack pipelines (BAP) that do not require internal knowledge of model architectures broadens the applicability and potential impact of adversarial attacks [24].

Assessing adversarial robustness involves evaluating the visual naturalness of adversarial attacks, which is critical for determining how well these attacks can blend into real-world environments without detection [33]. This evaluation is essential for understanding the practical implications of adversarial threats and the effectiveness of defense mechanisms.

Achieving adversarial robustness necessitates addressing intrinsic vulnerabilities that expose models to various adversarial threats, developing robust evaluation metrics to assess resilience, and implementing adaptive defense mechanisms to counteract a wide range of adversarial attacks, including those in digital and physical environments. This multifaceted approach must identify and mitigate specific attack vectors—such as model extraction, inversion, and poisoning attacks—while enhancing detection capabilities of convolutional neural networks (CNNs) against adversarial perturbations, ensuring reliability in real-world applications [34, 29, 14]. As adversarial techniques evolve, developing robust defenses remains a critical research area for the secure deployment of deep learning systems across diverse domains.

2.4 Real-World Applications and Adversarial Challenges

The deployment of deep learning systems in real-world applications faces adversarial challenges that threaten reliability and effectiveness. Object detection models, for instance, are significantly affected by environmental factors such as lighting and distance, which can diminish the efficacy of adversarial patches [35]. These variables complicate the deployment of adversarial defenses in dynamic settings where conditions are constantly changing.

AI-generated, for reference only.

In Traffic Sign Recognition (TSR) systems, the vulnerability to physical-world adversarial attacks is pronounced, particularly through hiding and appearing attacks that impair the detection of critical traffic signs [36]. The challenge is exacerbated by the inability of current adversarial methods to generate effective physical adversarial examples (AEs) in dynamic environments, where object positioning and video quality are in flux [37].

Adversarial patches, widely studied in digital contexts, face significant limitations in the physical world. Transformations such as rotations and shifts in 3D space can render 2D adversarial patches ineffective, necessitating more robust approaches that account for these transformations [38]. Furthermore, the requirement for detailed model knowledge in existing adversarial methods limits their practical applicability, as real-world scenarios often involve incomplete or obfuscated model information [39].

The use of adversarial attacks with laser beams exemplifies the diverse range of adversarial methods manifesting in real-world applications, underscoring the need for novel approaches to counter these threats [40]. Additionally, the gap in understanding physical backdoor attacks, as opposed to digital ones, presents a significant challenge in developing comprehensive defense strategies [41].

In autonomous driving scenarios, the PAN dataset provides valuable insights into the visual naturalness of adversarial attacks through human gaze data and Mean Opinion Scores (MOS), offering a benchmark for assessing the stealthiness of adversarial perturbations [33]. This evaluation is crucial for understanding how well adversarial attacks can blend into everyday environments without detection.

Innovative methods, such as synthesizing multiple views of 3D objects and applying adversarial perturbations, have been proposed to enhance model resistance to attacks, demonstrating the potential for self-supervised adversarial training to improve robustness [42]. These advancements are essential for ensuring the secure and reliable operation of deep learning systems across various real-world applications, including user activity sensing and heart rate detection in IoT-integrated environments [16]. As adversarial challenges continue to evolve, developing adaptive and resilient defense mechanisms remains a critical research area.

3 Adversarial Attacks in the Physical World

The emergence of physical adversarial attacks has garnered significant attention in machine learning, particularly due to their exploitation of vulnerabilities in deep learning systems in real-world settings. Unlike digital attacks, which operate under controlled conditions, physical adversarial attacks involve perturbations influenced by environmental factors like lighting and object deformation. This section explores various types of physical adversarial attacks, their mechanisms, effectiveness, and implications for system security.

To illustrate these concepts, Figure 2 presents a comprehensive overview of the hierarchical structure of physical adversarial attacks. This figure details the various types of attacks, the challenges encountered in their execution, and their implications for system security. Additionally, it highlights the methods employed in crafting these attacks, the impact they have on system performance, and provides real-world case studies that exemplify their application across different sectors. By integrating this visual representation, we can better understand the complexities and ramifications of physical adversarial attacks in contemporary machine learning environments.

3.1 Types of Physical Adversarial Attacks

Physical adversarial attacks introduce perturbations that interact with real-world conditions, distinguishing them from digital attacks. Adversarial patches, for instance, can impair real-time object detection by targeting superficial features in DNNs [33, 5]. However, these patches often lack visual consistency with their backgrounds, making them detectable and computationally demanding [21]. Despite challenges, adversarial patches remain effective despite physical transformations due to their stable positioning [23].

Innovative methods, such as using neon beams or laser spots, enhance the stealth and effectiveness of attacks. The AdvLS method uses laser spots as dynamic perturbations, optimized through genetic algorithms for covert daytime attacks [24]. Similarly, the Reflected Light Attack (RFLA) employs

AI-generated, for reference only.

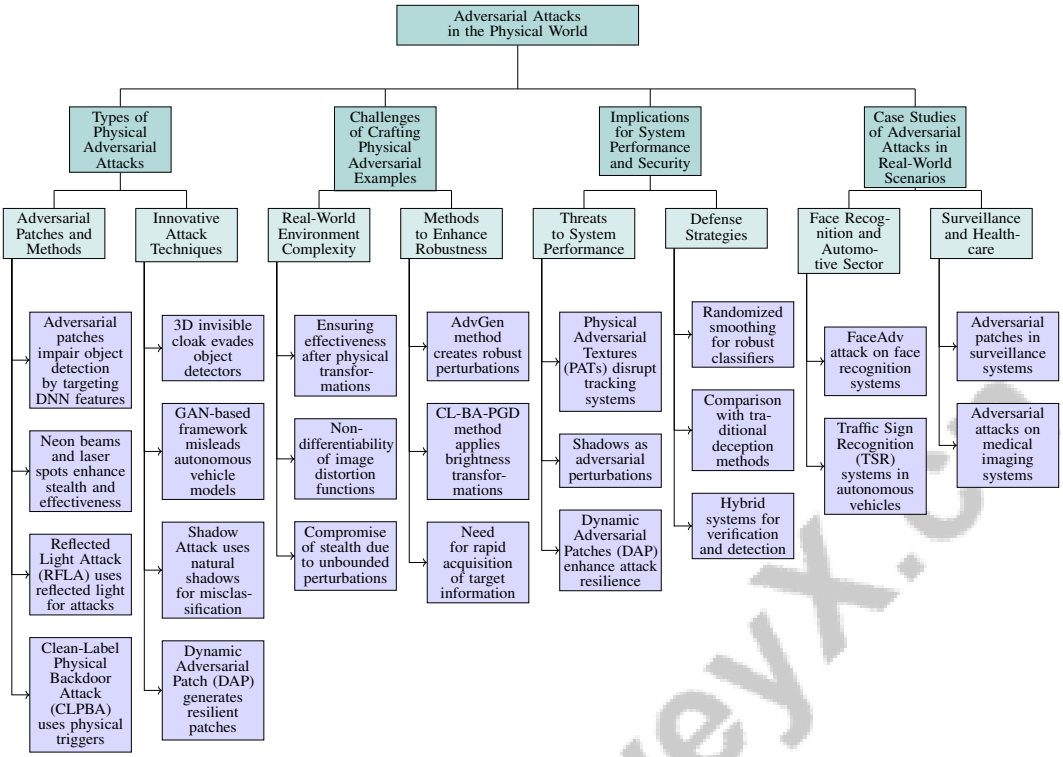


Figure 2: This figure illustrates the hierarchical structure of physical adversarial attacks, detailing their types, challenges, implications for system security, and real-world case studies. It highlights the methods used, challenges faced in crafting these attacks, their impact on system performance, and examples of their application in various sectors.

reflected light to create effective attacks in both digital and physical environments. Backdoor attacks pose significant threats in physical contexts, where varying conditions obscure triggers, leading to high failure rates for digital methods [19]. The Clean-Label Physical Backdoor Attack (CLPBA) uses physical triggers in training datasets while preserving original labels, showcasing sophisticated strategies in physical settings.

A 3D invisible cloak, achieved by printing adversarial patches on clothing, allows evasion from object detectors [17]. This underscores the need for robust defenses against evolving threats. Additionally, a GAN-based framework generates adversarial examples that mislead autonomous vehicle steering models at roadside signs, illustrating manipulation potential in critical systems.

Experiments reveal that many adversarial examples remain effective despite physical transformations, highlighting their robustness. However, transferring digital attacks to real-world applications is challenging, as physical adversarial patch attacks often fail to induce significant model errors. The Shadow Attack uses natural shadows to induce misclassification in machine learning models, achieving high success rates on LISA and GTSRB test sets while maintaining stealth [26, 43]. The Dynamic Adversarial Patch (DAP) generates natural-looking patches resilient to real-world distortions, while the BrPatch reduces visibility by manipulating brightness in the RGB color space. The AdvCL attack uses catoptric light properties to create adversarial perturbations that mislead DNNs.

Detection methods often focus on specific attacks and do not generalize well across different physical adversarial attacks [21]. Many methods emphasize digital adversarial perturbations, failing to represent real-world conditions accurately. The main challenge lies in the digital-physical-digital conversion process, which degrades adversarial examples' effectiveness due to variations in distance, angle, and illumination.

AI-generated, for reference only.

3.2 Challenges of Crafting Physical Adversarial Examples

Crafting effective physical adversarial examples is challenging due to the complex and dynamic nature of real-world environments. Ensuring effectiveness after exposure to physical transformations—such as lighting changes, perspective shifts, and object deformation—remains a significant hurdle. This arises from the non-differentiability of image distortion functions in real-world visual systems, which can alter adversarial perturbations' appearance. Existing methods often compromise stealth by allowing unbounded perturbations, resulting in noticeable artifacts. Recent advancements, including robust feature injection and 3D adversarial object synthesis, aim to create imperceptible adversarial examples effective across diverse conditions [44, 45, 26, 46, 47]. Traditional generation techniques frequently fail to retain adversarial properties under minor transformations, limiting their real-world applicability.

Robust adversarial examples require modeling physical transformations encountered in real-world scenarios. The AdvGen method exemplifies this by creating perturbations that remain effective after physical distortions, enhancing attack robustness [4]. Similarly, the CL-BA-PGD method addresses environmental variability by applying non-linear brightness transformations to training images, producing more resilient adversarial examples [48].

Another critical challenge is the rapid and precise acquisition of target information in dynamic environments. Existing non-contact optical attack methods often rely on time-consuming algorithms, limiting their efficiency in real-world applications [49]. This underscores the need for innovative techniques that adapt quickly to changing conditions without sacrificing effectiveness.

Benchmark evaluations of Vision and Language Models (VLAMs) reveal susceptibility to physical threats, including Out-of-Distribution (OOD) attacks, Typography-based Visual Prompts, and Adversarial Patch Attacks. These challenges highlight the necessity for adversarial examples to be robust against various physical-world perturbations for practical applicability [50].

Developing physical adversarial examples requires understanding critical factors influencing their effectiveness, including environmental variability, machine learning models' vulnerabilities, and the capacity to withstand physical transformations. This complexity is compounded by the unique characteristics of physical adversarial examples, affected by manufacturing processes and re-sampling techniques, necessitating exploration of robust strategies enhancing stealth and transferability while maintaining imperceptibility to human observers [51, 3, 47, 46]. Addressing these challenges is crucial for developing adversarial examples that reliably deceive models in real-world scenarios.

Figure 3 illustrates the primary challenges, innovative methods, and evaluation impacts associated with crafting physical adversarial examples. It highlights the complexity of ensuring robustness against physical transformations, environmental variability, and dynamic conditions, alongside advancements in adversarial methods and their evaluation through benchmarks.

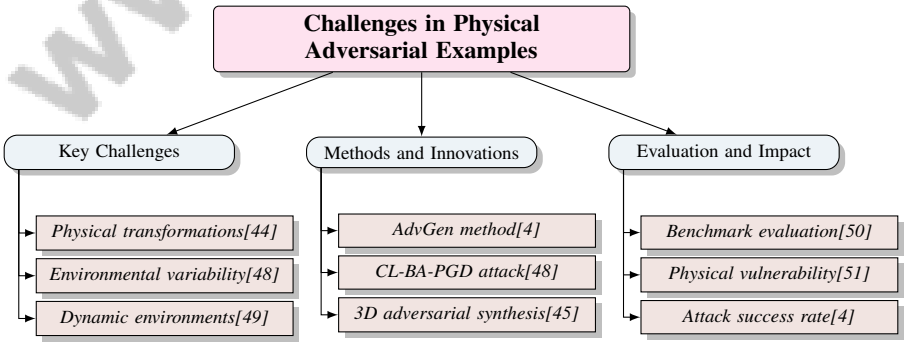


Figure 3: This figure illustrates the primary challenges, innovative methods, and evaluation impacts associated with crafting physical adversarial examples. It highlights the complexity of ensuring robustness against physical transformations, environmental variability, and dynamic conditions, alongside advancements in adversarial methods and their evaluation through benchmarks.

AI-generated, for reference only.

3.3 Implications for System Performance and Security

Adversarial attacks pose significant threats to the performance and security of deep learning systems, particularly in high-stakes contexts like autonomous driving, surveillance, and medical diagnostics. These attacks exploit model vulnerabilities, leading to incorrect outputs that severely compromise system integrity. For instance, Physical Adversarial Textures (PATs) disrupt tracking systems, illustrating adversarial attacks' profound impact on performance [17]. The ability of these attacks to induce errors underscores the necessity for robust defenses to maintain reliability.

Utilizing shadows as adversarial perturbations exemplifies the stealth of certain attack strategies, as shadows are natural and less likely to be detected compared to artificial modifications [26]. This method highlights the potential for everyday environmental features to serve as inconspicuous adversarial triggers, posing unique challenges to system security.

Dynamic Adversarial Patches (DAP) offer significant advantages over existing methods by producing less conspicuous patches resilient to real-world transformations [52]. Their resilience enhances effectiveness, complicating the defense landscape for deep learning systems.

Developing robust classifiers through techniques like randomized smoothing, which constructs certifiably robust classifiers by randomizing inputs, presents a promising avenue for enhancing system security [32]. This approach mitigates adversarial attacks' impact, ensuring models maintain performance amid perturbations.

Empirical evaluations emphasize the importance of comparing adversarial attacks with traditional non-ML-based deception methods to understand their relative impact on system performance and security [53]. This comparison is vital for developing comprehensive defense strategies addressing both traditional and adversarial threats.

The implications of adversarial attacks on deep learning systems' performance and security are significant and multifaceted, underscoring the urgent need for ongoing research and robust defenses. These attacks exploit inherent vulnerabilities, leading to severe misclassifications, particularly in critical applications like face recognition. Recent studies highlight various attack methods and their effectiveness, revealing that even imperceptible adversarial examples can have profound consequences. Researchers explore innovative strategies, including hybrid systems leveraging classical machine learning for verification and detection methodologies analyzing neural networks' unique activation patterns. This underscores the necessity for a proactive approach to fortify deep learning systems against evolving adversarial threats [29, 14, 34, 8, 54]. Ensuring secure and reliable operation of deep learning systems in real-world applications demands innovative solutions capable of countering the evolving landscape of adversarial threats.

3.4 Case Studies of Adversarial Attacks in Real-World Scenarios

Real-world case studies of adversarial attacks reveal their profound impact on deep learning systems' performance and security. The FaceAdv attack, for instance, demonstrated significant improvements in physical-world attack success rates on face recognition systems compared to existing methods [55]. This attack underscores face recognition systems' vulnerability to adversarial perturbations that bypass security mechanisms, potentially breaching identity verification processes.

In the automotive sector, research highlights the alarming susceptibility of Traffic Sign Recognition (TSR) systems in autonomous vehicles to physical-world adversarial attacks. These low-cost attacks can mislead TSR systems by obscuring critical traffic signs or presenting deceptive ones. Experimental results indicate such attacks can achieve a 100

A notable case study involves adversarial patches in surveillance systems, demonstrating how attackers can obscure identities or mislead detection algorithms within monitored environments. These patches exploit vulnerabilities in advanced object detection frameworks, such as YOLO, allowing individuals to evade recognition by surveillance cameras. Research underscores significant security implications, as these attacks can be executed in real-world scenarios using inconspicuous patterns on clothing or printed materials. Innovative detection methods have been proposed to counter these threats, focusing on both signature-based and signature-independent approaches to enhance surveillance resilience [25, 56, 57]. Such patches exploit object detection models' vulnerabilities, allowing adversaries to evade detection or manipulate surveillance footage, compromising security operations and enabling unauthorized activities.

AI-generated, for reference only.

In healthcare, adversarial attacks on medical imaging systems raise concerns about diagnostic accuracy and reliability. Subtle perturbations to medical images can alter diagnostic outcomes, potentially leading to incorrect treatment decisions. This highlights the necessity for enhancing adversarial robustness in medical applications to safeguard patient safety and maintain diagnostic accuracy amid emerging threats [46, 44, 58, 14].

These case studies illustrate the diverse range of applications affected by adversarial attacks and emphasize the need for continuous research and robust defense mechanisms. As adversarial techniques in deep learning advance, developing adaptive solutions that mitigate these threats while enhancing system robustness and security across various real-world applications is crucial. Addressing vulnerabilities—from imperceptible adversarial examples leading to misclassification to specific attack types like model extraction and poisoning—requires focused research on identifying weaknesses and creating effective countermeasures. Understanding adversaries' capabilities and goals, alongside the complexities of deploying robust defenses, is essential for ensuring reliable operation of deep learning technologies in increasingly complex environments [29, 14].

4 Evaluating Robustness Against Adversarial Attacks

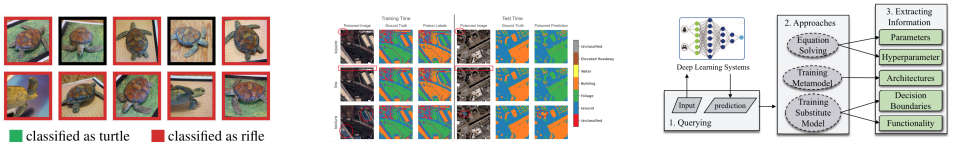
To effectively evaluate the robustness of deep learning systems against adversarial attacks, it is crucial to explore diverse methodologies and frameworks that facilitate comprehensive assessments. The following subsections delve into specific methodologies, frameworks, and metrics employed in robustness evaluation, highlighting their significance in addressing the complexities of adversarial threats.

4.1 Methodologies and Frameworks for Robustness Evaluation

Evaluating the robustness of deep learning systems against adversarial attacks requires diverse methodologies and frameworks to address the complexities of adversarial threats in real-world scenarios. The Expectation Over Transformation (EOT) framework constructs adversarial examples effective over a range of transformations, enhancing the realism and applicability of adversarial evaluations in physical environments [44]. The Physically-Realizable Adversarial Attack (PRAA) method creates inconspicuous adversarial patches optimized for effectiveness across multiple viewpoints, underscoring the importance of maintaining effectiveness despite environmental changes [23].

The TOUAP framework, evaluated with datasets like FLIR V2 1 and LLVIP, emphasizes comparing adversarial methods in both digital and physical environments for comprehensive robustness assessments [21]. In aerial detection contexts, experiments using DOTA and RSOD datasets demonstrate the effectiveness of contextual adversarial attacks, highlighting the need for robust evaluation methodologies that account for domain-specific challenges [5]. The PAN benchmark incorporates human perception and gaze data, enhancing the reliability of naturalness assessments [33].

The DC method employs strategies like SPONGE and BLINDING, showcasing the need for diverse evaluation strategies to assess the impact of different adversarial threats [19]. Black-box attack methods demonstrate the significance of evaluating adversarial robustness where model internals are inaccessible [24]. These methodologies and frameworks reflect the complexity of adversarial threats and the necessity for comprehensive assessment strategies to ensure secure and reliable operation across domains. As adversarial techniques evolve, developing robust evaluation frameworks is vital for assessing countermeasure effectiveness and ensuring deep learning algorithms withstand diverse attacks [44, 14].



(a) Classification of Objects: A Visual Comparison of Different Objects[59]

(b) Poisoned Image Detection in Satellite Imagery[60]

(c) Deep Learning Systems for Querying and Extracting Information[29]

Figure 4: Examples of Methodologies and Frameworks for Robustness Evaluation

AI-generated, for reference only.

As depicted in Figure 4, evaluating robustness against adversarial attacks is a crucial research area. The examples illustrate potential vulnerabilities in classification tasks, the detection of manipulated satellite images, and the processing of adversarial queries in deep learning systems [59, 60, 29].

4.2 Metrics for Robustness Evaluation

Robustness evaluation against adversarial attacks necessitates comprehensive metrics reflecting system performance under adversarial conditions. The Attack Success Rate (ASR) measures adversarial attack effectiveness by quantifying successful attack instances leading to incorrect outputs, identifying security vulnerabilities [54, 61, 29, 14]. Average Precision drop (AP drop) evaluates performance degradation in object detection systems due to adversarial attacks [62].

In Traffic Sign Recognition (TSR) systems, metrics like SysHA and SysAA analyze adversarial attack impacts on system performance, incorporating spatial memorization effects [36]. Evaluating adversarial example imperceptibility ensures perturbations blend into natural inputs without detection, maintaining effectiveness under various conditions [63].

Metrics like mean Average Precision (mAP), Naturalness Score (NS), and Transferability Score (TS) assess adversarial patch performance in effectiveness, visual naturalness, and cross-model transferability [64]. Additionally, metrics quantifying both benign performance and performance under attack provide a holistic view of model robustness [60]. Accuracy metrics compare defense methods against state-of-the-art data-end defenses in digital and physical scenarios, highlighting approach effectiveness [11].

These metrics are crucial for assessing deep learning system resilience, offering insights into vulnerabilities exploitable by adversarial attacks. By systematically analyzing these threats and their impact, researchers can develop robust models capable of handling diverse adversarial challenges in practical applications, enhancing system security and reliability [61, 29, 14].

4.3 Challenges in Evaluating Traditional Defenses

Evaluating traditional defenses against adversarial attacks in deep learning systems presents challenges due to model complexity and opacity. A significant issue is the lack of interpretability and systematic testing criteria, hindering effective defect detection and deployment in real-world applications [61]. The adaptability of attackers underscores the need for advanced evaluation metrics to assess defense resilience against diverse adversarial conditions, including model extraction, model inversion, poisoning, and physical attacks [54, 29, 14].

Traditional defenses often focus on specific attack types, limiting their generalizability against broader adversarial threats. Replicating real-world conditions during defense evaluation is challenging, as controlled lab settings fail to capture complexities found in practical applications, leading to quality issues like adversarial example generation [65, 1, 66, 30]. The lack of standardized benchmarks for assessing defense robustness complicates evaluation, emphasizing the need for systematic testing and improved metrics [29, 14, 66, 61, 67].

Developing comprehensive and interpretable evaluation frameworks is crucial for accurately assessing traditional defense robustness against diverse adversarial threats. Establishing metrics to evaluate countermeasure effectiveness and provide insights into deep learning model security weaknesses is essential [54, 29, 14].

4.4 Innovations in Robustness Evaluation Techniques

Recent advancements in robustness evaluation techniques for deep learning systems introduce novel methodologies addressing adversarial attack challenges in digital and physical environments. The Tight L0-norm Certified Robustness framework provides formal guarantees for model robustness using randomized smoothing techniques [32]. Self-supervised adversarial training techniques synthesize multiple 3D object views, enhancing model resilience against adversarial threats [42].

Benchmarking frameworks incorporating perceptual metrics, like the PAN dataset, advance robustness evaluation by considering human perception factors [33]. Two-stage optimized unified adversarial perturbations (TOUAP) improve robustness evaluation across visible and infrared spectrums [21].

AI-generated, for reference only.

Environmental feature exploitation, such as shadows and lighting, leads to innovative evaluation strategies accounting for subtle perturbations [26].

These advancements signify progress in identifying and addressing vulnerabilities exploited by adversarial threats. By adopting advanced methodologies, researchers can enhance model reliability and security in real-world applications, ensuring robust performance against evolving adversarial challenges [29, 14].

5 Enhancing Adversarial Robustness

Category	Feature	Method
Innovative Adversarial Attack Methods	Dynamic Patterns Efficiency Enhancement	BL[68], UPC[18] SFD[69], TAA[70]
Defense Mechanisms and Frameworks	Localized Defense Strategies	PG[10], AdvCL[71]
Techniques for Enhancing Robustness	Adversarial Defense Strategies Feature Selection Techniques Efficiency and Scalability	DAS[9], PRAA[23], PAT[17] RA[32] PDPA[72]
Developing Adaptive and Resilient Defense Mechanisms	Dynamic Defense Strategies	EVILEYE[73]

Table 1: This table provides a comprehensive summary of recent advancements in adversarial attack methods, defense mechanisms, and techniques for enhancing robustness in deep learning systems. It categorizes the methods into innovative adversarial attacks, defense frameworks, and strategies for increasing model resilience, highlighting key features and methodologies. The table underscores the evolving landscape of adversarial threats and the critical need for adaptive and resilient defense strategies in diverse applications.

Category	Feature	Method
Innovative Adversarial Attack Methods	Dynamic Patterns Efficiency Enhancement	BL[68], UPC[18] SFD[69], TAA[70]
Defense Mechanisms and Frameworks	Localized Defense Strategies	PG[10], AdvCL[71]
Techniques for Enhancing Robustness	Adversarial Defense Strategies Feature Selection Techniques Efficiency and Scalability	DAS[9], PRAA[23], PAT[17] RA[32] PDPA[72]
Developing Adaptive and Resilient Defense Mechanisms	Dynamic Defense Strategies	EVILEYE[73]

Table 2: This table provides a comprehensive summary of recent advancements in adversarial attack methods, defense mechanisms, and techniques for enhancing robustness in deep learning systems. It categorizes the methods into innovative adversarial attacks, defense frameworks, and strategies for increasing model resilience, highlighting key features and methodologies. The table underscores the evolving landscape of adversarial threats and the critical need for adaptive and resilient defense strategies in diverse applications.

To effectively enhance the adversarial robustness of deep learning systems, it is essential to first understand the nature and evolution of adversarial threats. Table 2 presents a detailed classification of contemporary adversarial attack methods and defense mechanisms, offering insight into the ongoing efforts to bolster the robustness of deep learning systems against sophisticated adversarial threats. Table 4 offers a comprehensive overview of innovative adversarial attack methods, illustrating their strategic approaches and adaptability, which underscore the necessity for enhanced defense mechanisms in deep learning systems. This understanding lays the groundwork for exploring innovative adversarial attack methods, which have become increasingly sophisticated and challenging for existing defense mechanisms. The subsequent subsection delves into these innovative attack strategies, highlighting their implications for the security and reliability of deep learning models in diverse applications.

5.1 Innovative Adversarial Attack Methods

Recent advancements in adversarial attack methodologies have introduced innovative techniques that pose significant challenges to the robustness of deep learning systems. Table 3 offers a comprehensive overview of innovative adversarial attack methods, illustrating their strategic approaches and adaptability, which underscore the necessity for enhanced defense mechanisms in deep learning systems. One such method is the development of BadLANE, a backdoor attack strategy that leverages amorphous trigger patterns. These patterns can be activated by various forms of environmental factors,

AI-generated, for reference only.

Method Name	Attack Strategies	Adaptability	Defense Implications
BL[68]	Backdoor Attacks	Dynamic Scene Adaptation	Advanced Defense Mechanisms
SFD[69]	Black-box Attacks	Adaptive Sampling Methods	Improve Defenses
TAA[70]	Targeted Perturbations	Different Scenarios	Robust Defenses
UPC[18]	Camouflage Patterns	Universal Camouflage Pattern	Advanced Defense Mechanisms

Table 3: This table provides a comparative analysis of various adversarial attack methods, highlighting their unique attack strategies, adaptability to different environments, and the implications for defense mechanisms. The methods discussed include backdoor attacks, black-box attacks, targeted perturbations, and universal camouflage patterns, each presenting distinct challenges to the robustness of deep learning systems.

such as mud spots or pollution, making them highly adaptable to dynamic scenes and increasing their resilience against traditional defenses [68]. This adaptability underscores the necessity for developing robust defense mechanisms that can effectively counteract the diverse triggers employed in such attacks.

In the realm of deep reinforcement learning (DRL), black-box attack strategies have been proposed to compromise models without requiring access to their internal parameters [69]. These attacks highlight the vulnerability of DRL systems to external manipulations and emphasize the need for defense strategies that can protect models against adversarial threats even in the absence of detailed model information.

The Targeted Physical-World Attention Attack (TAA) introduces a novel approach by utilizing soft attention maps to focus perturbations on critical pixels rather than applying uniform perturbations across the entire input space [70]. This targeted approach increases the efficiency and effectiveness of adversarial attacks, posing a significant challenge to existing defense mechanisms that are designed to counteract more generalized perturbations.

The Universal Physical Camouflage Attack (UPC) presents another innovative strategy by creating a universal pattern capable of deceiving multiple instances of the same object category. This attack targets the region proposal network, classifier, and regressor simultaneously, demonstrating the potential for universal patterns to disrupt the functionality of deep learning systems across various applications [18]. The implications of such attacks for defense strategies are profound, as they necessitate the development of more sophisticated detection and mitigation techniques that can address the universal nature of these adversarial patterns.

These novel adversarial attack methods highlight the evolving landscape of adversarial threats and the critical need for ongoing research to develop adaptive and resilient defense strategies. As adversarial techniques in machine learning continue to evolve, the need to enhance the robustness and security of deep learning systems against various forms of attacks—such as model extraction, poisoning, and adversarial examples—has become increasingly critical, particularly given the widespread deployment of these systems in high-stakes real-world applications where even minor vulnerabilities can lead to significant misclassification and security breaches. [1, 8, 29, 14]

5.2 Defense Mechanisms and Frameworks

The development of defense mechanisms and frameworks is essential for countering adversarial attacks in deep learning systems, particularly within Cyber-Physical Systems (CPS). These systems face unique security challenges due to their integration with the physical world, necessitating innovative defense strategies that address inherent vulnerabilities [10]. A notable approach is the PatchGuard framework, which employs Convolutional Neural Networks (CNNs) with small receptive fields and a robust masking technique to ensure provable robustness against localized adversarial patches [10]. This method exemplifies the potential for architectural innovations to contribute to the development of effective defense strategies.

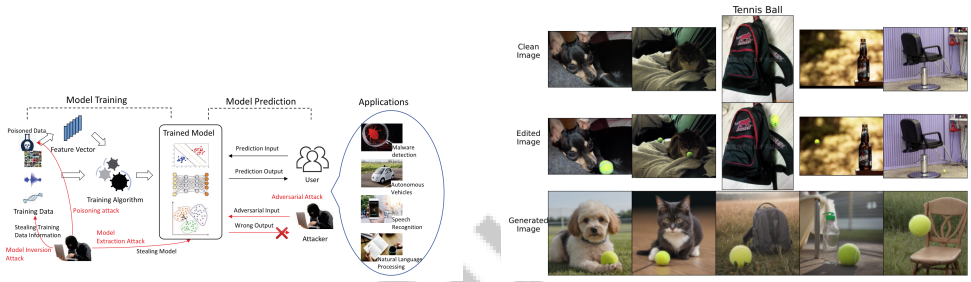
Adversarial training has been explored as a defense mechanism, particularly against adversarial catoptric light (AdvCL) attacks. While adversarial training can reduce the attack success rate (ASR), it does not completely neutralize the attack, indicating the need for complementary defense strategies [71]. This highlights the importance of developing multi-faceted defense mechanisms that can adapt to the evolving landscape of adversarial threats.

AI-generated, for reference only.

The security implications of backdoor attacks on object detection systems are particularly concerning, given their widespread use in daily life [20]. The understanding and addressing of these vulnerabilities are critical for enhancing the resilience of deep learning systems. Specialized defense mechanisms that focus on detecting and mitigating backdoor threats are necessary to safeguard the integrity of these systems.

Furthermore, the inadequacy of current defenses to counter physical backdoor attacks underscores the need for new strategies tailored to this specific domain. The prevalence of adversarial attacks on deep learning systems presents a substantial security challenge, highlighting the urgent need for the development of specialized defense mechanisms that can effectively counteract various attack types, such as model extraction, model inversion, poisoning, and adversarial perturbations, to minimize their detrimental effects on model performance and integrity. [29, 14, 34, 26, 74]. Understanding the vulnerabilities of backdoor attacks is critical for developing more resilient deep neural networks (DNNs) and improving the security of machine learning applications.

The ongoing advancement of adversarial attack techniques, which exploit the vulnerabilities inherent in deep learning systems, underscores the critical need for sustained research and the development of adaptive and robust defense strategies. These strategies are essential to safeguard the secure and reliable functioning of deep learning applications across diverse real-world scenarios, especially given the increasing sophistication of attacks that can mislead models through imperceptible perturbations or direct physical manipulations. [29, 14, 34, 8, 54]



(a) Poisoning Attacks on Machine Learning Models: A Comprehensive Overview[29]

(b) Editing a Tennis Ball Image[75]

Figure 5: Examples of Defense Mechanisms and Frameworks

As shown in Figure 5, In the realm of enhancing adversarial robustness, understanding the vulnerabilities of machine learning models to various attacks is crucial. The provided examples in Figure 5 illustrate two significant aspects of defense mechanisms and frameworks against adversarial threats. The first example, "Poisoning Attacks on Machine Learning Models: A Comprehensive Overview," offers a detailed visualization of how poisoning attacks can compromise the integrity of machine learning models. It breaks down the attack process into three critical phases: model training, model prediction, and applications, thereby emphasizing the necessity for robust training protocols to mitigate these threats. The second example, "Editing a Tennis Ball Image," creatively demonstrates how image manipulation can be employed to understand and test the resilience of models against physical backdoor attacks. By showcasing a series of edited images where a tennis ball is replaced with various objects, this example highlights the playful yet serious nature of adversarial testing, underscoring the importance of developing frameworks that can withstand such manipulations. Together, these examples serve as a foundation for exploring and enhancing defense mechanisms to bolster the robustness of machine learning systems against adversarial attacks. [**?**]he2020securitythreatsdeeplearning,yang2024synthesizingphysicalbackdoordatasets)

5.3 Techniques for Enhancing Robustness

Enhancing the robustness of deep learning systems against adversarial attacks requires the deployment of advanced techniques and strategies that improve model resilience to adversarial perturbations. One effective approach is the optimization of adversarial textures through guided adversarial losses and the Expectation Over Transformation (EOT) algorithm. This method ensures that adversarial

AI-generated, for reference only.

examples maintain their effectiveness across a range of transformations, thereby enhancing robustness in dynamic environments [17].

The Dual Attention Suppression (DAS) method exemplifies a strategy that enhances robustness by distracting model attention from target regions while ensuring high semantic correlation with the scenario context. This technique is particularly effective in maintaining model performance under adversarial conditions by redirecting attention away from potential perturbations [9].

Establishing reliable benchmarks for assessing visual naturalness is another crucial aspect of enhancing robustness. Such benchmarks provide a framework for evaluating the naturalness of adversarial examples, which is essential for developing more effective defenses in real-world scenarios [33]. This approach highlights the importance of evaluating models in realistic settings to ensure their robustness against adversarial threats.

The proposed method for physically realizable adversarial attacks significantly reduces navigation success rates by about 40

Randomized ablation is another technique that contributes to robustness by creating an ablated input through the random selection of a subset of features from the original input, setting the rest to a special value. This method enhances model resilience by ensuring that critical features are preserved even in the presence of adversarial perturbations [32].

The Parallelized Data Processing Algorithm (PDPA) introduces a novel approach to enhancing the speed and efficiency of handling large datasets by distributing data processing tasks across multiple cores. This technique is crucial for improving the robustness of models in handling large-scale data, particularly in edge computing environments [72].

The techniques and strategies discussed in the literature represent a multifaceted approach to bolstering the resilience of deep learning systems. This comprehensive framework not only addresses the inherent vulnerabilities—such as those exposed by model extraction, adversarial, and poisoning attacks—but also emphasizes the importance of understanding the learning tasks and their impact on model robustness. By integrating insights from various attack models and proposing targeted mitigations, these approaches ensure that deep learning systems can operate securely and reliably across a wide array of real-world applications, ultimately enhancing their performance in critical domains like image recognition, natural language processing, and more. [29, 14, 30, 76, 1]. As adversarial techniques continue to evolve, ongoing research and innovation remain critical to advancing the field of adversarial robustness.

5.4 Developing Adaptive and Resilient Defense Mechanisms

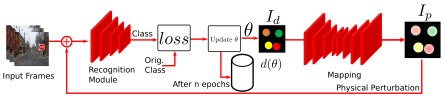
The development of adaptive and resilient defense mechanisms is crucial for ensuring the robustness of deep learning systems against evolving adversarial threats. As adversarial attack methodologies continue to advance, defense strategies must also evolve to effectively counter these sophisticated threats. A key focus of future research should be on developing defenses that can withstand perceptible attacks, which are designed to be noticeable yet still effective in misleading models [54]. These defenses should be capable of dynamically adapting to the changing landscape of adversarial threats, ensuring that models maintain their integrity and reliability in diverse real-world contexts.

Adaptive defense mechanisms should incorporate real-time monitoring and response capabilities, enabling systems to detect and mitigate adversarial attacks as they occur. To effectively address the challenges posed by adversarial perturbations in input data, it is essential to integrate sophisticated detection algorithms capable of recognizing subtle alterations, alongside the development of resilient response strategies designed to mitigate the impacts of these perturbations. This involves leveraging advanced methodologies such as randomized smoothing to enhance classifier robustness, as well as implementing comprehensive detection frameworks like DoPa, which analyze activation patterns in Convolutional Neural Networks (CNNs) to differentiate between adversarial and natural inputs. [1, 34, 32]. By leveraging machine learning techniques such as reinforcement learning, defense mechanisms can be trained to anticipate adversarial actions and adjust their strategies accordingly, enhancing their resilience to future attacks.

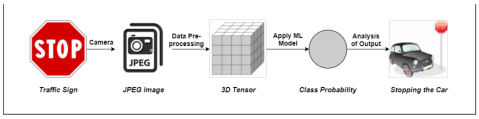
Moreover, the exploration of the implications of adversarial attacks in various real-world contexts is essential for developing comprehensive defense strategies. Understanding how different environmental factors and application scenarios influence the effectiveness of adversarial attacks can provide

AI-generated, for reference only.

valuable insights into the vulnerabilities of deep learning systems. This knowledge can guide the development of customized defense mechanisms specifically designed to counteract various types of adversarial attacks, such as model extraction and backdoor attacks, thereby enhancing the security and reliability of deep learning systems across diverse applications and threat scenarios. [75, 29, 14]



(a) A Deep Learning Framework for Physical Perturbation in Object Recognition[73]



(b) Traffic Sign Recognition System[14]

Figure 6: Examples of Developing Adaptive and Resilient Defense Mechanisms

As shown in Figure 6, In the realm of enhancing adversarial robustness, developing adaptive and resilient defense mechanisms is crucial for safeguarding machine learning systems from adversarial attacks. The examples illustrated in Figure 3 highlight two distinct yet complementary approaches to achieving this goal. The first example, "A Deep Learning Framework for Physical Perturbation in Object Recognition," showcases a sophisticated framework designed to handle physical perturbations in object recognition tasks. This framework integrates several components, including input frames, a recognition module, and a loss function, to iteratively update parameters and map physical perturbations, thereby improving the system's robustness against adversarial alterations. The second example, "Traffic Sign Recognition System," presents a flowchart detailing a system that begins with a camera capturing an image of a traffic sign. This image undergoes data preprocessing to form a 3D tensor, which is then processed by a machine learning model to output class probabilities. By analyzing these probabilities, the system determines the appropriate action, such as whether a vehicle should stop, thereby demonstrating a practical application of resilient defense mechanisms in real-world scenarios. Together, these examples underscore the importance of adaptive strategies in fortifying machine learning models against adversarial threats. [**?**]han2023dontcleanglassesperception,chakraborty2018adversarialattacksdefencessurvey)

Feature	BadLANE	Black-box Attack	Targeted Physical-World Attention Attack
Attack Strategy	Backdoor Attack	External Manipulation	Soft Attention Maps
Target Domain	Dynamic Scenes	Deep Reinforcement Learning	Physical World
Unique Feature	Amorphous Triggers	NO Internal Access	Critical Pixel Focus

Table 4: This table provides a comparative analysis of three innovative adversarial attack methods: BadLANE, Black-box Attack, and Targeted Physical-World Attention Attack. It highlights key features such as attack strategy, target domain, and unique characteristics, offering insights into their strategic approaches and implications for deep learning system robustness.

6 Real-World Applications and Case Studies

The integration of deep neural networks (DNNs) in real-world systems such as face recognition, object detection, and traffic sign recognition has introduced vulnerabilities that adversarial attacks can exploit. This section explores these vulnerabilities, particularly in face recognition and verification systems, underscoring the urgent need for robust defenses to ensure their integrity and reliability in practical applications.

6.1 Face Recognition and Verification Systems

Face recognition systems, powered by DNNs, have achieved remarkable accuracy in identification and verification but remain susceptible to adversarial attacks. These attacks exploit model weaknesses, leading to misclassification, particularly in physical environments [55]. Adversarial patches can deceive models by being discreetly placed on faces, posing significant threats to system integrity [77].

To counter these threats, robust defense mechanisms like Ad-YOLO have been developed, achieving an Average Precision (AP) of 80.31% against white-box attacks while maintaining detection accuracy

AI-generated, for reference only.

[78]. Backdoor attacks, using common objects as triggers, further highlight the need for defenses against sophisticated threats [79, 20].

Recent advancements in adversarial methodologies have produced techniques that generate effective adversarial examples in physical environments [80]. Evaluations of these methods in real-world applications, such as those from Google Play, reveal significant attack success rates, emphasizing the need for comprehensive defense strategies [24]. The adaptability of adversarial methods extends beyond face recognition, necessitating ongoing research to address challenges faced by FR systems in diverse applications [81].

6.2 Adversarial Attacks on Object Detection and Recognition

Adversarial attacks pose significant challenges to object detection and recognition systems, especially in real-world settings. Cross-modal adversarial patches have demonstrated high efficacy, targeting both visible and infrared detectors with an attack success rate exceeding 80% [62]. Adversarial camera patches (ADCP) further impair model performance, emphasizing the need for robust defenses against subtle perturbations [82].

Frameworks like DTA develop robust camouflage patterns that evade detection models in both simulated and real-world settings [83]. However, empirical evaluations suggest that some physical adversarial patch attacks may be overstated, as they can be ineffective against certain models [53]. The Universal Physical Camouflage (UPC) attack exemplifies a highly effective approach, necessitating sophisticated defenses against universal adversarial patterns [18].

The impact of adversarial attacks on object detection systems necessitates continuous research and development of robust defenses. Innovative solutions must address various adversarial threats, including model extraction, inversion, poisoning, and attacks exploiting inherent vulnerabilities [34, 29, 14].

6.3 Traffic Sign Recognition and Autonomous Driving

Adversarial robustness is crucial for traffic sign recognition (TSR) and autonomous driving systems, where accurate road sign interpretation is essential for safety. These systems are vulnerable to adversarial attacks, risking misclassification and erroneous actions. The DCI dataset offers valuable insights for evaluating TSR and autonomous driving under various conditions [31].

Techniques like creating shadows on traffic signs highlight the importance of adversarial robustness in TSR applications [26]. Defensive patch generation frameworks have improved model robustness, achieving over 20% accuracy enhancement against adversarial and corruption robustness [11]. TOUAP's effectiveness in deceiving detectors across environments underscores the need for robust defenses [21].

The Detector Collapse (DC) method represents a significant advancement in backdoor attack methodologies, underscoring the necessity for robust defenses to protect TSR and autonomous driving systems [19]. Recent findings expose vulnerabilities in autonomous driving systems, emphasizing the urgent need for robust defenses to maintain integrity and safety [37, 84, 14, 85].

6.4 Adversarial Robustness in Security and Surveillance

Adversarial robustness is paramount in security and surveillance systems, where model reliability is crucial for operational integrity. Techniques like AdvCF demonstrate the sophistication of adversarial threats across various conditions [86]. Benchmarks for evaluating physical-world adversarial attacks on TSR systems are essential for advancing robustness [36].

In near-infrared (NIR) based AI models, establishing benchmarks highlights critical security implications, underscoring the need for tailored defense strategies [87]. Techniques like VRAP illustrate significant threats to DNN-enabled applications, emphasizing the necessity for robust defense mechanisms [88].

Advancements in anomaly detection frameworks, achieving high F1-scores, illustrate potential for identifying attacks within Cyber-Physical Systems [89]. The exploration of imperceptible adversarial

AI-generated, for reference only.

attack methods presents novel challenges, necessitating innovative strategies to ensure security and reliability [90].

The significance of adversarial robustness in security systems cannot be overstated. As adversarial techniques evolve, developing adaptive and resilient defense mechanisms is essential to address vulnerabilities exposed by adversarial examples, ensuring security and reliability [54, 14].

6.5 Applications in Communication and Signal Processing

The integration of deep learning into communication and signal processing enhances waveform recognition and data security but introduces adversarial challenges. Inconspicuous adversarial patches (IAP) highlight the potential for attacks to exploit vulnerabilities while minimizing detection [57]. The Low Interception Waveform (LIW) method exemplifies advancements in mitigating adversarial risks, enhancing security [91].

Adversarial examples in radio frequency communications emphasize the implications of these threats, necessitating sophisticated defenses for reliable operation [92]. Digital twin technology in beam management offers promising solutions, enhancing communication processes [93].

The synthesis of physical backdoor datasets illustrates the potential for attacks to exploit communication system vulnerabilities [75]. Efficient data processing techniques, like the Parallelized Data Processing Algorithm (PDPA), enhance communication system robustness [72].

Continuous research is essential to develop robust countermeasures, ensuring the security and effectiveness of deep learning applications amidst rapid advancements and sophisticated threats [1, 29, 14, 30]. As adversarial techniques evolve, developing adaptive and resilient defense mechanisms remains critical to safeguarding these systems against an ever-changing landscape of threats.

7 Challenges and Future Directions

7.1 Limitations of Current Defensive Methods

Current defensive strategies for deep learning systems face significant limitations, particularly when addressing adversarial attacks in real-world settings. The effectiveness of adversarial textures is often inconsistent, influenced by environmental conditions and specific model implementations, necessitating adaptable defense strategies [17]. The complexity of training on extensive datasets, coupled with the black-box nature of models and challenges in unsupervised learning, further complicates defense efforts [1]. This underscores the need for transparent and interpretable defenses capable of handling complex data.

Contextual attack methods often rely on specific features, which may not always be prominent, highlighting the need for versatile defenses [5]. Additionally, assumptions about 2D image boards for view synthesis can lead to inaccuracies in real-world modeling [7]. The reliance of adversarial methods on specific trigger patterns limits their effectiveness in dynamic environments, emphasizing the need for adaptive defenses [19].

Existing benchmarks for evaluating adversarial examples often rely on subjective assessments, introducing bias [33]. Objective frameworks are crucial for accurate assessments. The applicability of certain methods is restricted in obfuscated environments due to difficulties in identifying model patterns [24]. Lastly, some defenses struggle under specific conditions or with certain object types, indicating a need for comprehensive approaches [23].

These limitations highlight the urgent need for innovative strategies to bolster the robustness and security of deep learning systems against evolving adversarial threats. Despite some countermeasures, many remain ineffective against new techniques exploiting existing vulnerabilities. A comprehensive understanding of these weaknesses and the effectiveness of proposed defenses is essential to advance the resilience of deep learning systems in the face of increasingly sophisticated adversarial challenges [94, 29, 14]. Developing adaptable defense strategies is crucial for ensuring reliable operation in real-world applications.

AI-generated, for reference only.

7.2 Future Research Directions

Future research should prioritize developing defenses against physical backdoor attacks, focusing on minimizing false positives in real-world applications [41]. This includes enhancing trigger specificity and reducing false activations through improved adversarial training techniques [95]. Exploring the implications of false positives across applications remains critical.

Refining encoding strategies and examining the impact of additional learning tasks on robustness are promising avenues for enhancing model resilience [76]. Research should also aim to develop robust defense mechanisms against physical attacks and explore the application of existing benchmarks in other domains [96]. Understanding the vulnerabilities of DNNs and developing countermeasures for object detection systems are crucial steps forward [97].

Emerging trends suggest focusing on enhancing verification algorithms and exploring design principles for more verifiable neural networks [98]. This could involve refining neural network models and investigating unsupervised learning for improved interpretability and robustness [99].

In CPS, research should expand the capabilities of large language models (LLMs), improve prompt design, and integrate expert knowledge for better task performance [16]. Developing frameworks for responsible AI use, exploring AI as a Service (AIaaS), and addressing model reliability and fairness challenges are critical areas [100].

Research should also refine the input generation process for autonomous driving systems, enhance the load-balancing mechanism of the Parallelized Data Processing Algorithm (PDPA), and explore its application in different domains [72]. Investigating multi-sensor fusion attacks and the impact of dynamic adversarial vehicles will advance defenses to enhance autonomous driving safety [2].

Moreover, future research should enhance adversarial attack robustness in diverse settings and develop sophisticated evaluation metrics [25]. Optimizing hue mapping methods and enhancing the stealthiness of adversarial patches are also important [101]. Further exploration into the real-world application of the Dual Attention Suppression (DAS) method will significantly contribute to the field [9].

Additionally, improving the robustness of adversarial patches against countermeasures and testing them in diverse scenarios are critical [23]. Enhancing the robustness of Physical Adversarial Textures (PATs) and optimizing textures for black-box models are further study areas [17]. Finally, research should focus on enhancing model interpretability, developing robust unsupervised learning methods, and exploring cross-modal learning to improve adaptability across data types [1].

7.3 Broader Implications and Future Directions

Adversarial robustness has far-reaching implications across domains, emphasizing the need for secure and reliable deep learning systems. As attacks evolve, their impact on sectors like autonomous driving, healthcare, and infrastructure becomes significant. The integration of LLMs with IoT systems highlights the expanding role of AI, underscoring the need for robust defenses [16].

Future research should focus on adaptive defense mechanisms that dynamically respond to evolving threats. This includes exploring reinforcement learning to anticipate adversarial actions and adjust strategies in real-time, enhancing resilience. Refining evaluation frameworks to incorporate perceptual metrics, as demonstrated by the PAN dataset, offers valuable insights into the visual naturalness of attacks, guiding effective defense strategies [33].

Exploring adversarial robustness in CPS presents opportunities to address AI integration challenges with physical infrastructure. Future research should expand LLM capabilities in CPS tasks, improve prompt design, and integrate expert knowledge [16]. Developing frameworks for responsible AI use and exploring AIaaS are critical for ethical AI deployment [100].

In autonomous driving, refining input generation processes and exploring dynamic safety configurations are essential [85]. Investigating multi-sensor fusion attacks and the impact of dynamic adversarial vehicles will advance defenses for autonomous systems [2].

These broader implications underscore the importance of continued research and innovation in developing adaptive defense mechanisms, refining evaluation frameworks, and exploring ethical AI deployment. By systematically addressing inherent vulnerabilities—such as model extraction,

AI-generated, for reference only.

inversion, poisoning, and adversarial attacks—researchers can enhance security and reliability, ensuring effective deployment across various applications. This comprehensive approach mitigates threats and fosters robust frameworks capable of withstanding sophisticated attacks [1, 29, 14, 30].

7.4 Understanding Cyber-Physical Interactions and Environmental Variability

Cyber-physical interactions introduce complexities impacting adversarial robustness in deep learning models. CPS are susceptible to attacks due to dynamic environments. Environmental variability, such as lighting and weather changes, affects adversarial example performance, challenging model robustness in real-world scenarios [17].

This variability is pronounced in autonomous driving and traffic sign recognition, where real-time decision-making is crucial. Attacks exploiting environmental factors—like shadows—can deceive models with perturbations indistinguishable from natural variations [26]. These perturbations necessitate robust defenses maintaining performance across diverse conditions.

Cyber-physical interactions also pose unique challenges. Integrating AI with physical infrastructure requires processing data from various sensors, each subject to noise and interference. Attacks targeting specific sensors, like LiDAR, can induce errors [2].

To tackle these challenges, adaptive defense mechanisms must dynamically respond to variability, detect diverse attacks, and account for cyber-physical interactions. This approach should leverage deep learning advancements and address threats identified in recent research, ensuring robust protection against evolving tactics and enhancing overall security [73, 29, 102, 14]. Techniques like multi-sensor fusion can enhance resilience against threats by integrating data from multiple sources for a comprehensive understanding, improving resistance to perturbations.

Advancing adversarial robustness requires comprehensive evaluation frameworks incorporating environmental variability, as it significantly influences performance and security against examples. These frameworks should address uncertainties in deep learning decisions, considering real-world conditions affecting adversarial example generation and detection, enhancing reliability in safety-critical domains [66, 103, 14]. They should include metrics assessing real-world condition impacts on performance, providing insights into defense effectiveness.

Understanding cyber-physical interactions and environmental variability's impact on robustness is crucial for secure, reliable deep learning systems in real applications. By addressing inherent weaknesses and threats—like adversarial attacks, model extraction, and poisoning—researchers can develop resilient models better equipped to navigate dynamic environments, enhancing reliability and effectiveness [1, 66, 29, 30].

7.5 Transferability and Generalization of Adversarial Examples

The transferability of adversarial examples, their ability to deceive multiple models, poses a significant threat to deep learning systems' security across applications [24]. This generalization complicates defenses, indicating vulnerabilities extending beyond individual models.

Understanding transferability challenges involves considering model differences, such as architecture and training data, which influence susceptibility to attacks [4]. Environmental variability, like lighting changes, affects generalization in real-world scenarios [17].

To address these challenges, robust evaluation frameworks assessing transferability and generalization across models and environments are essential. These should include metrics evaluating adversarial example impacts, providing insights into transferability and generalization capabilities. Exploring ensemble methods, combining predictions of multiple models, can enhance resilience by reducing the likelihood of all models being deceived by the same perturbation [32].

Understanding mechanisms contributing to transferability and generalization is crucial for effective defense strategies. Investigating model decision boundaries and feature representations helps facilitate transferability. By understanding attack mechanisms, researchers can develop resilient models mitigating vulnerabilities, reducing misclassification likelihood. This enables designing systems that withstand manipulation and incorporate verification processes, like integrating classical machine learning models for enhanced security [1, 8, 29, 14].

AI-generated, for reference only.

Addressing transferability and generalization challenges is critical for secure, reliable deep learning systems across applications. By developing comprehensive evaluation frameworks assessing system vulnerabilities and exploring innovative defense strategies, researchers can enhance resilience against sophisticated threats. This approach addresses inherent weaknesses and lays the groundwork for effective countermeasures adapting to new methodologies, improving AI applications' security and reliability [94, 29, 14].

7.6 Evaluation and Benchmarking of Robustness

Benchmark	Size	Domain	Task Format	Metric
PAN[33]	2,688	Autonomous Driving	No-Reference Image Quality Assessment (nr-IQA)	SROCC, PLCC
PBDB[75]	1,000,000	Computer Vision	Backdoor Attack Simulation	Attack Success Rate, Clean Accuracy
SDC-Benchmark[12]	12,000	Lane Keeping	Behavioral Cloning	Steering Angle, Lateral Deviation
DRP[84]	33,000	Lane Detection	Adversarial Attack Simulation	Success Rate, Average Success Time
LLM-VLM-Robotics[104]	150	Robotics	Manipulation	Success Rate, Input Similarity
TSR-Bench[36]	1,000	Traffic Sign Recognition	Adversarial Attack Evaluation	SysHA, SysAA
NIR-ADA[87]	13,000	Surveillance	Human Detection	Average Confidence, Attack Success Rate
Drive-By-Fly-By[105]	32,000	Traffic Sign Detection	Object Detection	Attack Success Rate

Table 5: This table presents a comprehensive overview of various benchmarks used to evaluate the robustness of deep learning systems against adversarial attacks. It includes key attributes such as benchmark size, domain, task format, and evaluation metrics, providing insights into the diverse scenarios and metrics employed in robustness assessments.

Evaluating and benchmarking deep learning systems' robustness against adversarial attacks is crucial for understanding vulnerabilities and enhancing resilience. Robustness evaluation provides insights into performance under adversarial conditions, guiding secure system development. Evaluation identifies model weaknesses and establishes benchmarks for comparing defense strategies [33].

Benchmarking frameworks, like the PAN dataset, offer comprehensive assessments incorporating human perception and gaze data, enhancing naturalness assessments' reliability [33]. This benchmark is critical for understanding adversarial robustness's perceptual aspects, ensuring evaluations consider human factors. The PAN dataset provides insights into adversarial perturbations' visual naturalness, guiding effective defense strategy development.

Standardized benchmarks provide consistent bases for evaluating model and defense mechanism robustness. These should incorporate diverse adversarial scenarios, including digital and physical attacks, ensuring comprehensive performance assessments. Including diverse attack types, like patches, backdoor attacks, and environmental perturbations, is crucial for understanding adversarial threats' full spectrum and their impact on robustness [17].

Integrating perceptual metrics into evaluations offers nuanced understanding of adversarial examples' perception in real scenarios. Considering factors like visual naturalness and attack stealthiness, researchers can develop effective defense mechanisms addressing technical and perceptual robustness aspects [33].

Evaluating and benchmarking robustness are essential for advancing adversarial robustness in deep learning systems, helping identify and mitigate vulnerabilities to various attack types, enhance countermeasure effectiveness, and ensure models maintain performance under real conditions and transformations [29, 14, 44, 63, 59]. By providing a structured framework for assessing performance under adversarial conditions, researchers can identify vulnerabilities, develop resilient models, and ensure secure, reliable operation across diverse applications. Table 5 provides a detailed overview of representative benchmarks utilized for assessing the robustness of deep learning systems, highlighting their domains, task formats, and evaluation metrics.

AI-generated, for reference only.

8 Conclusion

Enhancing adversarial robustness in deep learning systems is crucial for their secure deployment in real-world applications where security and reliability are imperative. This survey highlights the persistent vulnerability of these models to adversarial attacks that exploit both digital and physical domains to induce incorrect outputs. The creation of robust adversarial examples accentuates the practical risks associated with these threats. Recent progress in defense strategies, such as Jujutsu and LanCe, showcases their effectiveness in countering adversarial threats while minimizing false positives, reflecting the necessity for continual innovation in adversarial defense techniques. Additionally, self-supervised adversarial training has shown promise in bolstering the robustness of Monocular Depth Estimation models against diverse attacks, underscoring the potential of adaptive learning methods. The exploration of Large Language Models for interpreting physical signals opens avenues for integrating AI into cyber-physical systems, emphasizing the need for further research in this area. A comprehensive understanding of the underlying physical principles of deep learning models is vital for crafting effective defenses. The survey underscores the importance of sustained research and development to enhance the adversarial robustness of deep learning systems. As adversarial methods evolve, ensuring these systems' secure and dependable operation in real-world contexts remains a formidable challenge, necessitating innovative approaches and collaborative efforts within the research community.

www.SurveyX.cn

AI-generated, for reference only.

References

[1] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[2] Yang Lou, Yi Zhu, Qun Song, Rui Tan, Chunming Qiao, Wei-Bin Lee, and Jianping Wang. A first physical-world trajectory prediction attack via lidar-induced deceptions in autonomous driving, 2024.

[3] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[4] Sai Amrit Patnaik, Shivali Chansoriya, Anil K. Jain, and Anoop M. Namboodiri. Advgen: Physical adversarial attack on face presentation attack detection systems, 2023.

[5] Jiawei Lian, Xiaofei Wang, Yuru Su, Mingyang Ma, and Shaohui Mei. Contextual adversarial attack against aerial detection in the physical world, 2023.

[6] Xingxing Wei, Ying Guo, and Jie Yu. Adversarial sticker: A stealthy attack method in the physical world, 2022.

[7] Zhiyuan Cheng, James Liang, Guanhong Tao, Dongfang Liu, and Xiangyu Zhang. Adversarial training of self-supervised monocular depth estimation against physical-world attacks, 2023.

[8] Mohammed Alkhowaiter, Hisham Kholidy, Mnassar Alyami, Abdulmajeed Alghamdi, and Cliff Zou. Adversarial-aware deep learning system based on a secondary classical machine learning verification approach, 2023.

[9] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world, 2021.

[10] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking, 2021.

[11] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. Defensive patches for robust recognition in the physical world, 2022.

[12] Andrea Stocco, Brian Pulfer, and Paolo Tonella. Mind the gap! a study on the transferability of virtual vs physical-world testing of autonomous driving systems, 2022.

[13] Camilo Pestana, Wei Liu, David Glance, Robyn Owens, and Ajmal Mian. Physical world assistive signals for deep neural network classifiers – neither defense nor attack, 2021.

[14] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018.

[15] Bao Gia Doan, Minhui Xue, Shiqing Ma, Ehsan Abbasnejad, and Damith C. Ranasinghe. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems, 2022.

[16] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. Penetrative ai: Making llms comprehend the physical world, 2024.

[17] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking, 2019.

[18] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors, 2020.

[19] Hangtao Zhang, Shengshan Hu, Yichen Wang, Leo Yu Zhang, Ziqi Zhou, Xianlong Wang, Yanjun Zhang, and Chao Chen. Detector collapse: Physical-world backdooring object detection to catastrophic overload or blindness in autonomous driving, 2024.

AI-generated, for reference only.

www.SlideX.cn

[20] Hua Ma, Yinshan Li, Yansong Gao, Alsharif Abuadbba, Zhi Zhang, Anmin Fu, Hyoungshick Kim, Said F. Al-Sarawi, Nepal Surya, and Derek Abbott. Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world, 2022.

[21] Chengyin Hu and Weiwen Shi. Two-stage optimized unified adversarial patch for attacking visible-infrared cross-modal detectors in the physical world, 2023.

[22] Takami Sato and Qi Alfred Chen. On robustness of lane detection models to physical-world adversarial attacks in autonomous driving, 2021.

[23] Meng Chen, Jiawei Tu, Chao Qi, Yonghao Dang, Feng Zhou, Wei Wei, and Jianqin Yin. Towards physically realizable adversarial attacks in embodied vision navigation, 2025.

[24] Hongchen Cao, Shuai Li, Yuming Zhou, Ming Fan, Xuejiao Zhao, and Yutian Tang. Towards black-box attacks on deep learning apps, 2021.

[25] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors, 2020.

[26] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon, 2022.

[27] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[28] Wei Jiang, Tianyuan Zhang, Shuangcheng Liu, Weiyu Ji, Zichao Zhang, and Gang Xiao. Exploring the physical world adversarial robustness of vehicle detection, 2023.

[29] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards security threats of deep learning systems: A survey, 2020.

[30] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[31] Tianyuan Zhang, Yisong Xiao, Xiaoya Zhang, Hao Li, and Lu Wang. Benchmarking the physical-world adversarial robustness of vehicle detection, 2023.

[32] Jinyuan Jia, Binghui Wang, Xiaoyu Cao, Hongbin Liu, and Neil Zhenqiang Gong. Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations, 2022.

[33] Simin Li, Shuing Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks, 2023.

[34] Zirui Xu, Fuxun Yu, and Xiang Chen. Dopa: A comprehensive cnn detection methodology against physical adversarial attacks, 2019.

[35] Jakob Shack, Katarina Petrovic, and Olga Saukh. Breaking the illusion: Real-world challenges for adversarial patches in object detection, 2024.

[36] Ningfei Wang, Shaoyuan Xie, Takami Sato, Yunpeng Luo, Kaidi Xu, and Qi Alfred Chen. Revisiting physical-world adversarial attack on traffic sign recognition: A commercial systems perspective, 2024.

[37] Wei Jia, Zhaojun Lu, Haichun Zhang, Zhenglin Liu, Jie Wang, and Gang Qu. Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems, 2022.

[38] Yi Wang, Jingyang Zhou, Tianlong Chen, Sijia Liu, Shiyu Chang, Chandrajit Bajaj, and Zhangyang Wang. Can 3d adversarial logos cloak humans?, 2020.

[39] Bo Luo and Qiang Xu. Region-wise attack: On efficient generation of robust physical adversarial examples, 2020.

AI-generated, for reference only.

www.SurveyX.cn

[40] Ranjie Duan, Xiaofeng Mao, A. K. Qin, Yun Yang, Yuefeng Chen, Shaokai Ye, and Yuan He. Adversarial laser beam: Effective physical-world attack to dnns in a blink, 2021.

[41] Emily Wenger, Josephine Passananti, Arjun Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world, 2021.

[42] Zhiyuan Cheng, Cheng Han, James Liang, Qifan Wang, Xiangyu Zhang, and Dongfang Liu. Self-supervised adversarial training of monocular depth estimation against physical-world attacks, 2024.

[43] Pedram MohajerAnsari, Alkim Domeke, Jan de Voor, Arkajyoti Mitra, Grace Johnson, Amir Salarpour, Habeeb Olufowobi, Mohammad Hamad, and Mert D. Pesé. Discovering new shadow patterns for black-box attacks on lane detection of autonomous vehicles, 2024.

[44] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018.

[45] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space, 2019.

[46] Yichen Wang, Yuxuan Chou, Ziqi Zhou, Hangtao Zhang, Wei Wan, Shengshan Hu, and Minghui Li. Breaking barriers in physical-world adversarial examples: Improving robustness and transferability via robust feature, 2024.

[47] Weilin Xu, Sebastian Szyller, Cory Cornelius, Luis Murillo Rojas, Marius Arvinte, Alvaro Velasquez, Jason Martin, and Nageen Himayat. Imperceptible adversarial examples in the physical world, 2024.

[48] Inderjeet Singh, Satoru Momiyama, Kazuya Kakizaki, and Toshinori Araki. On brightness agnostic adversarial examples against face recognition systems, 2021.

[49] Yitong Sun, Yao Huang, and Xingxing Wei. Embodied laser attack:leveraging scene priors to achieve agent-based robust non-contact attacks, 2024.

[50] Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, and Renjing Xu. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models, 2024.

[51] Jiakai Wang, Xianglong Liu, Jin Hu, Donghua Wang, Siyang Wu, Tingsong Jiang, Yuanfang Guo, Aishan Liu, and Jiantao Zhou. Adversarial examples in the physical world: A survey, 2024.

[52] Amira Guesmi, Ruitian Ding, Muhammad Abdullah Hanif, Ihsen Alouani, and Muhammad Shafique. Dap: A dynamic adversarial patch for evading person detectors, 2023.

[53] Gavin S. Hartnett, Li Ang Zhang, Caolionn O'Connell, Andrew J. Lohn, and Jair Aguirre. Empirical evaluation of physical adversarial patch attacks against overhead object detection models, 2022.

[54] Yuxin Cao, Yumeng Zhu, Derui Wang, Sheng Wen, Minhui Xue, Jin Lu, and Hao Ge. Rethinking the threat and accessibility of adversarial attacks against face recognition systems, 2024.

[55] Meng Shen, Hao Yu, Liehuang Zhu, Ke Xu, Qi Li, and Xiaojiang Du. Robust attacks on deep learning face recognition in the physical world, 2020.

[56] Bin Liang, Jiachun Li, and Jianjun Huang. We can always catch you: Detecting adversarial patched objects with or without signature, 2021.

[57] Tao Bai, Jinqi Luo, and Jun Zhao. Inconspicuous adversarial patches for fooling image recognition systems on mobile devices, 2021.

[58] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. X-adv: Physical adversarial object attacks against x-ray prohibited item detection, 2023.

AI-generated, for reference only.

www.SurveyX.cn

[59] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[60] Elise Bishoff, Charles Godfrey, Myles McKay, and Eleanor Byler. Quantifying the robustness of deep multispectral segmentation models against natural perturbations and data poisoning, 2023.

[61] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, et al. Deepgauge: Multi-granularity testing criteria for deep learning systems. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, pages 120–131, 2018.

[62] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world, 2023.

[63] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks, 2018.

[64] Zheng Zhou, Hongbo Zhao, Ju Liu, Qiaosheng Zhang, Liwei Geng, Shuchang Lyu, and Wenquan Feng. Mvpatch: More vivid patch for adversarial camouflaged attacks on object detectors in the physical world, 2024.

[65] Jinhan Kim, Robert Feldt, and Shin Yoo. Guiding deep learning system testing using surprise adequacy. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 1039–1049. IEEE, 2019.

[66] Xiyue Zhang, Xiaofei Xie, Lei Ma, Xiaoning Du, Qiang Hu, Yang Liu, Jianjun Zhao, and Meng Sun. Towards characterizing adversarial defects of deep learning software from the lens of uncertainty, 2020.

[67] Jinhan Kim, Robert Feldt, and Shin Yoo. Guiding deep learning system testing using surprise adequacy, 2018.

[68] Xinwei Zhang, Aishan Liu, Tianyuan Zhang, Siyuan Liang, and Xianglong Liu. Towards robust physical-world backdoor attacks on lane detection, 2024.

[69] Xinlei Pan, Chaowei Xiao, Warren He, Shuang Yang, Jian Peng, Mingjie Sun, Jinfeng Yi, Zijiang Yang, Mingyan Liu, Bo Li, and Dawn Song. Characterizing attacks on deep reinforcement learning, 2022.

[70] Xinghao Yang, Weifeng Liu, Shengli Zhang, Wei Liu, and Dacheng Tao. Targeted physical-world attention attack on deep learning models in road sign recognition, 2021.

[71] Chengyin Hu and Weiwen Shi. Adversarial catoptric light: An effective, stealthy and robust physical-world attack to dnns, 2023.

[72] Terence Jie Chua, Wenhan Yu, and Jun Zhao. Mobile edge adversarial detection for digital twinning to the metaverse with deep reinforcement learning, 2023.

[73] Yi Han, Matthew Chan, Eric Wengrowski, Zhuohuan Li, Nils Ole Tippenhauer, Mani Srivastava, Saman Zonouz, and Luis Garcia. Why don't you clean your glasses? perception attacks with dynamic optical perturbations, 2023.

[74] Zirui Xu, Fuxun Yu, and Xiang Chen. Lance: A comprehensive and lightweight cnn defense methodology against physical adversarial attacks on embedded multimedia applications, 2019.

[75] Sze Jue Yang, Chinh D. La, Quang H. Nguyen, Kok-Seng Wong, Anh Tuan Tran, Chee Seng Chan, and Khoa D. Doan. Synthesizing physical backdoor datasets: An automated framework leveraging deep generative models, 2024.

[76] Keji Han, Yun Li, Xianzhong Long, and Yao Ge. Learning task-aware robust deep learning systems, 2021.

AI-generated, for reference only.

www.SurveyX.cn

[77] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models, 2021.

[78] Nan Ji, YanFei Feng, Haidong Xie, Xueshuang Xiang, and Naijin Liu. Adversarial yolo: Defense human detection patch attacks via detecting adversarial patches, 2021.

[79] Xianlong Wang, Hewen Pan, Hangtao Zhang, Minghui Li, Shengshan Hu, Ziqi Zhou, Lulu Xue, Peijin Guo, Yichen Wang, Wei Wan, Aishan Liu, and Leo Yu Zhang. Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation, 2025.

[80] Inderjeet Singh, Toshinori Araki, and Kazuya Kakizaki. Powerful physical adversarial examples against practical face recognition systems, 2022.

[81] Xiao Yang, Yinpeng Dong, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Controllable evaluation and generation of physical adversarial patch on face recognition, 2022.

[82] Kalibinuer Tiliwalidi. Adversarial camera patch: An effective and robust physical-world attack on object detectors, 2023.

[83] Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network, 2022.

[84] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jack Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack, 2021.

[85] Ziwen Wan, Junjie Shen, Jalen Chuang, Xin Xia, Joshua Garcia, Jiaqi Ma, and Qi Alfred Chen. Too afraid to drive: Systematic discovery of semantic dos vulnerability in autonomous driving planning under physical-world attacks, 2022.

[86] Chengyin Hu and Weiwen Shi. Adversarial color film: Effective physical-world attack to dnns, 2023.

[87] Muyao Niu, Zhuoxiao Li, Yifan Zhan, Huy H. Nguyen, Isao Echizen, and Yinqiang Zheng. Physics-based adversarial attack on near-infrared human detector for nighttime surveillance camera systems, 2024.

[88] Xiaosen Wang and Kunyu Wang. Generating visually realistic adversarial patch, 2023.

[89] Nuno Oliveira, Norberto Sousa, Jorge Oliveira, and Isabel Praça. Anomaly detection in cyber-physical systems: Reconstruction of a prediction error feature space, 2021.

[90] Zvi Stein and Adrian Stern. Imperceptible cmos camera dazzle for adversarial attacks on deep neural networks, 2023.

[91] Haidong Xie, Jia Tan, Xiaoying Zhang, Nan Ji, Haihua Liao, Zuguo Yu, Xueshuang Xiang, and Naijin Liu. Low-interception waveform: To prevent the recognition of spectrum waveform modulation via adversarial examples, 2022.

[92] Silvija Kokalj-Filipovic, Rob Miller, Nicholas Chang, and Chi Leung Lau. Mitigation of adversarial examples in rf deep classifiers utilizing autoencoder pre-training, 2019.

[93] Shuaifeng Jiang and Ahmed Alkhateeb. Digital twin based beam prediction: Can we train in the digital world and deploy in reality?, 2023.

[94] Dang Duy Thang and Toshihiro Matsui. Search space of adversarial perturbations against image filters, 2020.

[95] Thinh Dao, Cuong Chi Le, Khoa D Doan, and Kok-Seng Wong. Towards clean-label backdoor attacks in the physical world, 2024.

AI-generated, for reference only.

www.Surveyx.cn

[96] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models, 2018.

[97] Mingfu Xue, Can He, Zhiyu Wu, Jian Wang, Zhe Liu, and Weiqiang Liu. 3d invisible cloak, 2020.

[98] Lindsey Kuper, Guy Katz, Justin Gottschlich, Kyle Julian, Clark Barrett, and Mykel Kochenderfer. Toward scalable verification for safety-critical deep networks, 2018.

[99] Dian Lei, Xiaoxiao Chen, and Jianfei Zhao. Opening the black box of deep learning, 2018.

[100] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[101] Mingzhen Shao. Brightness-restricted adversarial attack patch, 2023.

[102] Zhiyuan Yu, Zack Kaplan, Qiben Yan, and Ning Zhang. Security and privacy in the emerging cyber-physical world: A survey, 2021.

[103] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, 2019.

[104] Xiyang Wu, Souradip Chakraborty, Ruiqi Xian, Jing Liang, Tianrui Guan, Fuxiao Liu, Brian M. Sadler, Dinesh Manocha, and Amrit Singh Bedi. Highlighting the safety concerns of deploying llms/vlms in robotics, 2024.

[105] Bao Gia Doan, Dang Quang Nguyen, Callum Lindquist, Paul Montague, Tamas Abraham, Olivier De Vel, Seyit Camtepe, Salil S. Kanhere, Ehsan Abbasnejad, and Damith C. Ranasinghe. On the credibility of backdoor attacks against object detectors in the physical world, 2024.

www.SurveyX.cn

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

www.SurveyX.cn

AI-generated, for reference only.