Artificial Intelligence Review (2025) 58:373
https://doi.org/10.1007/s10462-025-11373-w

Securing (vision-based) autonomous systems: taxonomy, challenges, and defense mechanisms against adversarial threats

Alvaro Lopez Pellicer[1] · **Plamen Angelov**[1] · **Neeraj Suri**[1]

Received: 24 March 2025 / Accepted: 26 August 2025

Abstract
The rapid integration of computer vision into Autonomous Systems (AS) has introduced new vulnerabilities, particularly in the form of adversarial threats capable of manipulating perception and control modules. While multiple surveys have addressed adversarial robustness in deep learning, few have systematically analyzed how these threats manifest across the full stack and life-cycle of AS. This review bridges that gap by presenting a structured synthesis that spans both, foundational vision-centric literature and recent AS-specific advances, with focus on digital and physical threat vectors. We introduce a unified framework mapping adversarial threats across the AS stack and life-cycle, supported by three novel analytical matrices: the *Life-cycle–Attack Matrix* (linking attacks to data, training, and inference stages), the *Stack–Threat Matrix* (localizing vulnerabilities throughout the autonomy stack), and the *Exposure–Impact Matrix* (connecting attack exposure to AI design vulnerabilities and operational consequences). Drawing on these models, we define holistic requirements for effective AS defenses and critically appraise the current landscape of adversarial robustness. Finally, we propose the *AS-ADS* scoring framework to enable comparative assessment of defense methods in terms of their alignment with the practical needs of AS, and outline actionable directions for advancing the robustness of vision-based autonomous systems.

Keywords Artificial intelligence · Autonomous systems · Security · Computer vision · Adversarial attacks · Adversarial defenses

✉ Alvaro Lopez Pellicer
a.lopezpellicer@lancaster.ac.uk

Plamen Angelov
p.angelov@lancaster.ac.uk

Neeraj Suri
neeraj.suri@lancaster.ac.uk

[1] School of Computing and Communications, Lancaster University, InfoLab21, Lancaster LA1 4WA, UK

Published online: 08 October 2025

1 Introduction

Autonomous Systems (AS) are rapidly transitioning from research prototypes to mission-critical platforms in transportation, logistics, and robotics (Sheridan 2016; Siciliano and Khatib 2016). At their core, AS combine high-resolution sensors, fast communication links, complex control software, and deep neural networks to enable autonomous operation in unstructured environments (Bekey 2005; Guizzo 2011).

A defining trend in modern AS is their deep reliance on computer vision. Vision models, ranging from classic convolutional neural networks (CNNs) (He et al. 2016), real-time detectors such as YOLO (Wang et al. 2023a) and RT-DETR (Zhao et al. 2024a), to advanced transformers (Oquab et al. 2024) and vision-language models (Xu et al. 2024; Renz et al. 2024) underpin not only perception but also sensor fusion, semantic mapping, prediction, planning, and even direct actuation. The industry-wide move towards vision-centric and even vision-only paradigms is perhaps best exemplified by Tesla's Autopilot and Full Self-Driving systems (Tesla Inc 2022), which intentionally omit LiDAR and radar in favor of multi-camera, deep learning pipelines for end-to-end environment understanding and control.

While classic non-vision attack vectors such as GPS spoofing (Horton and Ranganathan 2018), CAN-bus injection (Kang et al. 2021), and physical attacks on radar or LiDAR systems (Cao et al. 2019; Kong et al. 2020) have been extensively studied, and industry best practices for their detection and mitigation are relatively mature, the shift to vision-centric architectures introduces a new class of system-wide vulnerabilities. Years of adversarial machine learning research have shown that even digital imperceptible perturbations to image inputs can induce misclassification and dangerous misinterpretation (Szegedy et al. 2013; Goodfellow et al. 2014). In the physical world, attacks such as adversarial stickers on traffic signs (Eykholt et al. 2018) or adversarial patches (Zhang et al. 2022a) among others demonstrate the persistence and transferability of adversarial threats across architectures and conditions.

Crucially, in vision-centric AS, a single vulnerability in perception rarely remains isolated. Because perception outputs directly feed into planning, prediction, and control with limited or no human oversight, adversarial effects can propagate, be amplified by sensor fusion or trajectory optimization, and ultimately result in system-level failures. This risk is heightened by the industry trend towards closed-loop, end-to-end architectures, where raw vision inputs may directly dictate vehicle or robot behavior.

Unlike static computer vision systems, AS operate in dynamic, multi-agent, and safety-critical environments (Bojarski et al. 2016; Janai et al. 2020). Attacks can target any phase, from data acquisition and model training to online operation or inter-vehicle communication, and their impact can extend far beyond classification accuracy, undermining safety, trust, and real-world performance in ways rarely captured by static benchmarks.

This review is motivated by the urgent need to understand adversarial vulnerabilities and defenses for vision-centric AS, bridging insights from both foundational adversarial machine learning and the fast-evolving AS-specific literature. By systematically mapping how threats propagate across the AS stack and life-cycle, we clarify real deployment challenges, highlight the limitations of existing approaches, and provide a unified analytical foundation for evaluating adversarial robustness in AS. Our survey intentionally bridges

the gap between the mature vision-centric adversarial ML literature and the recent but fast-growing AS-specific corpus.

1.1 Related work

Several recent surveys have addressed elements of adversarial attacks and defenses, but none provide a life-cycle-integrated, stack-specific analysis tailored to real-world AS. For example, Badjie et al. (2024) present a systematic review of adversarial attacks and countermeasures in image classification models for autonomous driving, with detailed coverage of attack types and proactive/reactive defenses. However, their analysis is limited to perception modules and does not examine attack propagation through planning and control subsystems, nor does it offer a unified threat model for the entire AS life-cycle. Akhtar et al. (2021), a comprehensive review of advances in adversarial attacks and defenses for computer vision is provided, focusing on algorithmic and architectural aspects after 2018. However, their work does not account for the layered structure or operational context of AS, omitting issues such as temporal vulnerability, subsystem coupling, or deployment-specific constraints.

Deng et al. (2021) provide a detailed analysis of different attacks and defenses in the workflow of the autonomous driving system, covering adversarial attacks for various deep learning models and attacks in both physical and cyber contexts. While comprehensive in scope, their survey does not offer a structured framework for evaluating defense strategies across different stages of the AS life-cycle. Liu et al. (2021) examine adversarial attacks and defenses from an interpretation perspective, providing valuable insight into model vulnerability, but focusing less on system-level threats specific to autonomous systems.

Almutairi and Barnawi (2023) present an overview of adversarial attacks, defenses, and frameworks to secure DNNs in smart vehicles, organizing their analysis around security challenges but lacking a cohesive approach to understanding cross-layer vulnerabilities. Similarly, Khamaiseh et al. (2022) provide an extensive survey on adversarial attacks and defense mechanisms for image classification, though their focus remains primarily on algorithmic approaches rather than on the operational contexts of autonomous systems.

Amirkhani et al. (2023) review prominent attack and defense mechanisms for object detection in autonomous vehicles, offering discussions on their strengths and weaknesses, but without addressing the integrated nature of attack surfaces across the entire autonomous vehicle stack. Boltachev (2024) highlights key types of disruptive attacks on autonomous driving models, demonstrating potential threats through experimental validation but not providing a systematic framework for defense evaluation.

Ibrahum et al. (2024) perform a systematic review of adversarial attacks and defenses in autonomous vehicles, prioritizing safety and introducing a taxonomy inspired by SOTIF. However, their focusis on risk scenarios and lacks an analytical framework linking attack surfaces, layered vulnerabilities, and defense evaluation across the AS stack. Girdhar et al. (2023) offer a review centered on cybersecurity in autonomous vehicles, highlighting known attack vectors and defenses but stopping short of providing an actionable structure for mapping attacks or evaluating defenses in an integrated, system-aware fashion.

Xu et al. (2020) broaden the perspective to attacks and defenses in images, graphs, and text, but their survey remains modality-driven and does not tackle the architectural and temporal challenges unique to AS. The work by Costa et al. (2024) surveys adversarial attacks

and defenses across various deep learning architectures, offering a high-level synthesis without focusing on the operational realities, threat models, or deployment constraints of AS. Malik et al. (2024) present a systematic review of adversarial machine learning attacks and defensive controls, but their analysis lacks the specificity required for autonomous systems operating in dynamic environments.

1.2 List of contributions

In contrast, our survey bridges the foundational adversarial machine learning concepts presented in Akhtar et al. (2021); Xu et al. (2020); Costa et al. (2024); Liu et al. (2021); Amirkhani et al. (2023); Malik et al. (2024); Khamaiseh et al. (2022) and the overly component-specialized AS surveys in Badjie et al. (2024); Ibrahum et al. (2024); Girdhar et al. (2023); Deng et al. (2021); Almutairi and Barnawi (2023); Boltachev (2024) with a holistic, layered systems analysis of AS, organized around **three key contributions:**

1. **Bridging gaps in existing surveys:** While prior reviews often isolate general adversarial ML or AS-specific applications, our work integrates foundational adversarial concepts, vision-based robustness literature, and AS-specific challenges into a unified analytical framework. This enables life-cycle-integrated thinking and supports the development of practical AS defenses.
2. **System-level threat modeling via analytical matrices:** We construct three matrices that connect existing adversarial literature to the specific vulnerabilities of AS:

    - The **Life-cycle–attack matrix** categorizes threats across the Data, Training, and Inference stages of the AI life-cycle, linking attack types (e.g., poisoning, backdoors, evasion) to stage-specific weaknesses and highlighting temporal exposure windows, (Sect. 4.1).
    - The **Exposure–impact matrix** organizes threats by AI design vulnerabilities (e.g., data hunger, model sensitivity), attack surfaces, and downstream consequences such as sabotage or system misguidance, providing a framework to understand full-system threat pathways in real-world AS deployments, (Sect. 4.2).
    - The **Stack–threat matrix** maps how adversarial attacks impact AS subsystems' Perception, Planning, and Control layers, demonstrating how vulnerabilities propagate and compound across the stack. We ground our analysis with realistic subsystem scenarios, target models, and operational implications, (Sect. 4.3).
    - Additionally, we provide a comparative synthesis of both **digital and physical adversarial attacks**, characterizing representative methods in terms of attack type, robustness, and practical implications. This serves as a unified reference for evaluating attack feasibility and severity in both real-world and simulation contexts, (Sect. 3).

Rather than serving as abstract taxonomies, these matrices function as actionable threat modeling tools to guide robustness benchmarking and inform future research.

3. **Critical appraisal and evaluation of defense strategies:** We develop a structured methodology to assess how well existing adversarial defenses meet the unique needs of AS:

   - Drawing from the literature and the threat matrices developed in this review, we derive a high-level set of **overall requirements** that adversarial defenses must satisfy to be viable in AS environments. Focusing on real-time constraints, adaptability, interpretability, and efficiency, (Sect. 5.1).
   - We examine the current landscape of defenses targeting **physical-world attacks**, identifying the strengths and limitations of existing approaches and clarifying where critical gaps remain, (Sect. 5.2).
   - We consolidate and simplify prior defense taxonomies, aligning them with AS-specific criteria to enable more meaningful evaluation across mechanism types, (Sect. 5.3).
   - Based on this foundation, we introduce the **Autonomous systems adversarial defense score (AS-ADS)**, a novel evaluation framework that scores defense methods across four deployment-relevant axes: *real-time capability*, *adaptability to novel threats*, *interpretability*, and *resource efficiency*, (Sect. 5.4).
   - To demonstrate the AS-ADS framework, we evaluate a representative subsample of 30 defense methods; 15 from the general vision adversarial robustness literature, and 15 from AS-specific works, highlighting the trade-offs and readiness of each, (Table 9):

This review, to the best of our knowledge, is the first to systematically bridge foundational adversarial machine learning and AS-specific literature in a holistic, layered systems analysis of Autonomous Systems.

1.3 Methodology and review protocol

This review implements a structured, reproducible literature survey based on PRISMA 2020 principles, specifically adapted to the context of machine learning and AS. Our goal is to comprehensively synthesize advances in adversarial robustness for vision-based models relevant to AS, bridging both foundational vision-centric theory and recent AS-specific developments.

We included works ranging from foundational studies (dating back to 1988) to the most recent publications available as of May 2025, identified through five major databases: IEEE Xplore, SpringerLink, ACM Digital Library, ScienceDirect, and arXiv (tracks: `cs.CV`, `cs.RO`, `stat.ML`). Search queries combined terms such as "adversarial attack," "defense," "autonomous systems," "dataset," "computer vision," "robotics," "LiDAR," and related phrases. After deduplication, non-vision and unrelated tracks were filtered, followed by manual screening of titles and abstracts. Full-text eligibility required methodological clarity, empirical evaluation, and relevance to either adversarial computer vision or AS.

Inclusion criteria were: (i) peer-reviewed venue (CORE A*/A/B or Scimago Q1–Q3 journal) or high-impact arXiv preprint, (ii) empirical focus on adversarial robustness, and (iii) coverage of vision models, pipelines, or AS-specific systems. Studies outside these domains, lacking empirical grounding, or duplicating prior work were excluded. Flexible

Table 1 Summary of the PRISMA screening resutls

Initial records identified	1041
Duplicates removed	99
Titles and abstracts screened	942
Excluded during abstract screening	614
Full-text articles assessed	328
Excluded after full-text review	91
Studies included in the final synthesis	237

Table 2 Breakdown of included papers by domain (vision-centric or AS-specific), era (foundational or recent), and contribution type (defense, attack, dataset, other). Percentages reflect the share of each row total

Domain	Era	Defense	Attack	Dataset	Other	Row Total
Vision-centric	Foundational (pre-2020)	39 (44.8%)	28 (32.2%)	3 (3.4%)	17 (19.5%)	87
Vision-centric	Non-foundational (2020+)	43 (60.6%)	21 (29.6%)	3 (4.2%)	4 (5.6%)	71
AS-specific	Foundational (pre-2020)	1 (6.7%)	2 (13.3%)	0 (0.0%)	12 (80.0%)	15
AS-specific	Non-foundational (2020+)	32 (50.0%)	17 (26.6%)	4 (6.3%)	11 (17.2%)	64
Column totals		115 (48.5%)	68 (28.7%)	10 (4.2%)	44 (18.6%)	237

Defense: Proposes, benchmarks, or surveys robustness mechanisms. *Attack:* Proposes, benchmarks, or surveys adversarial threats

Dataset: Introduces or is primarily a dataset/benchmark paper. *Other:* Surveys, theoretical, sensor, or general background works

inclusion criteria were applied to physical attack/defense and real-world system studies, reflecting their practical significance.

Following this protocol, we included **237 papers** in the final synthesis. Each was classified in a reproducible two-level taxonomy: (1) *Domain* (vision-centric or AS-specific), and (2) *Contribution Type* (defense, attack, dataset, or other supportive/background). Within each domain, references were further split as *foundational* (pre-2020) or *non-foundational* (2020 onward). Contribution types were assigned using a combination of keyword analysis (title/abstract), citation context (appearance in attack or defense tables/sections), and manual review for ambiguous cases. The domain split (vision-centric vs AS-specific) was established via systematic keyword matching and manual inspection for works with cross-domain relevance. While every effort was made to ensure comprehensive and reproducible coverage, we acknowledge the potential for misclassification in ambiguous cases and invite community feedback for future updates.

The review process and screening outcomes are summarized in Table 1.

Table 2 summarizes the final distribution of included studies by domain, era, and contribution type, supporting full reproducibility and transparency.

2 Background

Understanding adversarial robustness in AS requires grounding in the specific architectures, vision model deployments, and operational realities that distinguish AS from conventional computer vision systems. In practice, modern AS tightly integrate vision models not only for perception, but also across sensor fusion, prediction, planning, and closed-loop control, resulting in complex pathways for attack propagation and defense. The threat landscape in AS is shaped by this interconnectedness, exposing weaknesses that are rarely visible in

static, perception-only or digital-only evaluations. The limitations of current benchmarks and defense taxonomies, (most of which are tailored to standard image tasks), underscore the need for analysis methods and robustness criteria explicitly aligned with AS operational stacks and environment. This section provides the technical foundations, empirical context, and critical gaps necessary for our analysis.

2.1 Vision models & the autonomous system stack

Modern AS are fundamentally vision-driven, with deep learning models tightly integrated across nearly every functional layer; from perception to planning, control, and actuation. Unlike traditional computer vision pipelines, where outputs often remain within isolated modules, AS architectures are defined by close interconnection: the output of one model (e.g., object detection, segmentation) serves as direct input to downstream planning and control components, with minimal human oversight or redundancy.

The AS stack can be broadly divided into three groups: the **Physical Environment**, the **Hardware Layer**, and the **Hardware and Software Integration** layer, as shown in Fig. 1. The physical environment refers to the operational context, such as roadways for driverless vehicles or warehouse floors for robots. In the hardware layer we find sensors such as cameras (Forsyth and Ponce 2011; Szeliski 2022), LiDAR (Besl 1988; Hsu 2002), radar (Knee 2005; Hao et al. 2002), and ultrasonic sensors (Kinsler et al. 2000), which are often fused for greater robustness (Yeong et al. 2021) (sesor fusion). Communication hardware enables inter-device connectivity for federated learning (Yang et al. 2021), remote operations (Yu et al. 2021), or mission planning via satellite links (Prevot et al. 2016). Actuators close the hardware loop by translating digital commands into real-world action.

Across all layers, the adoption of general-purpose vision models, such as ResNet-50 (He et al. 2016), ViT (Dosovitskiy et al. 2020), SAM (Kirillov et al. 2023), and DINOv2 (Oquab

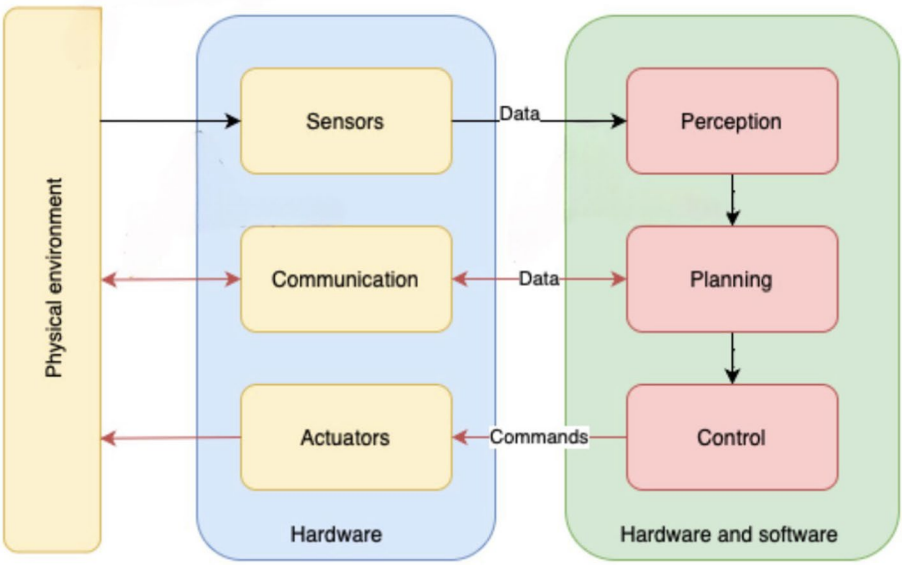

Fig. 1 Autonomous system stack diagram

et al. 2024), reflects the field's inheritance of both the strengths and adversarial vulnerabilities discovered in conventional computer vision. Specialized models (e.g., DriveVLM Tian et al. 2024, CarLLaVA Renz et al. 2024, BEVFormer Li et al. 2022) further illustrate the trend toward unified, stack-spanning pipelines.

More in depth, the AS **Perception layer** is now dominated by a broad spectrum of vision models. Ranging from early CNN backbones like ResNet-50 (He et al. 2016) to advanced architectures for detection and segmentation. Real-time detectors, such as YOLOv4 (Bochkovskiy et al. 2020), YOLOv7 (Wang et al. 2023a), RT-DETR (Zhao et al. 2024a), and EfficientDet (Tan et al. 2020), enable high-throughput object and obstacle identification. For segmentation and spatial reasoning, models like DeepLabv3+ (Chen et al. 2018a), Mask R-CNN (He et al. 2017), and SAM (Kirillov et al. 2023) provide fine-grained environmental parsing, while ViT (Dosovitskiy et al. 2020) and DINOv2 (Oquab et al. 2024) represent the adoption of transformer-based and foundation models. Multi-modal sensor fusion architectures–DAIR-V2X (Zhao et al. 2024b), UMoE (Lou et al. 2023), COMPASS (Ma et al. 2022) integrate camera, LiDAR, and other modalities for richer world models. Classical two-stage detectors like Fast R-CNN (Girshick 2015), Faster R-CNN (Ren et al. 2015), SSD (Liu et al. 2016a), and RetinaNet (Lin et al. 2017) also persist in specific AS deployments.

Within the **Planning layer**, outputs from perception are translated into actionable decisions and trajectories using a new wave of context-aware models. BEVFormer (Li et al. 2022) performs multi-view, spatiotemporal fusion for 3D scene understanding. Vision-language models such as DriveVLM (Tian et al. 2024), CarLLaVA (Renz et al. 2024), and VLM-AD (Xu et al. 2024) incorporate semantic context and agent interaction for robust closed-loop planning. End-to-end pipelines such as DAVE2 (Bojarski et al. 2016), PilotNet (Bojarski et al. 2017), and Conditional Imitation Learning (Codevilla et al. 2018) map visual or multimodal input directly to navigation actions, bypassing rule-based intermediaries. Legacy approaches such as ChauffeurNet (Bansal et al. 2018) and ALVINN (Pomerleau 1988) laid the groundwork for behavior prediction and direct perception-control mapping.

At the **Control layer** AS increasingly embed neural controllers, building upon foundations like ALVINN (Pomerleau 1988) towards deep reinforcement and imitation learning models (Lillicrap et al. 2015; Pan et al. 2017; Dursun et al. 2025), to execute planned actions in real time. These controllers handle adaptive actuation, closed-loop correction, and safe responses to unstructured or adversarial environments. Classic rule-based and PID controllers are now frequently augmented or replaced by neural networks that leverage features from vision and planning models for fine-grained actuation, error recovery, and robust operation under uncertainty. This integration enables rapid, flexible adjustment, but also exposes the system to error propagation: a perturbation at perception or planning can now directly alter low-level control, amplifying the risk of system-level failures.

Because the AS stack is tightly coupled and feedback-driven, whether at the sensor interface, within fusion modules, or at the control output, vulnerabilities in vision models cannot be isolated locally. Perturbations at any point in the stack can cascade through planning and control, ultimately triggering unexpected or catastrophic outcomes. This architecture demands adversarial robustness methods that are not only perception-aware, but explicitly stack and life-cycle-aware as well. A central principle developed throughout this review.

2.2 Adversarial threats in autonomous systems

The concept of *adversarial examples* was first introduced in Szegedy et al. (2013), who showed that deep learning models can be deceived by carefully crafted, human-imperceptible perturbations to input data. Formally, adversarial attacks seek to modify a given input $x_0 \in \mathbb{R}^d$ to a new point $x \in \mathbb{R}^d$, such that x is assigned a specific target class by the model, differing from the original prediction. The perturbation $\delta = x - x_0$ is typically constrained to be small in a chosen norm (e.g., $\|\delta\|_p < \epsilon$) to ensure that x remains visually indistinguishable from x_0 to humans. Methods such as *evasion attacks* employ optimization techniques, including the box-constrained L-BFGS algorithm (Fletcher 2013), to compute minimal perturbations that induce misclassification. Notably, these adversarial examples are often *transferable*. A single perturbation generated for one model can also mislead other deep neural networks–raising serious concerns for the security and reliability of AI systems as originally demonstrated in Liu et al. (2016b); Papernot et al. (2016a).

In the context of AS, **digital attacks** (e.g., FGSM Goodfellow et al. 2014, PGD Madry et al. 2019, C&W Carlini and Wagner 2017b) remain important, operating at inference or training time to introduce pixel-level perturbations or backdoors (e.g., BadNets Gu et al. 2017, MetaPoison Huang et al. 2020). These attacks, originally evaluated on canonical datasets like ImageNet or CIFAR, have proven highly transferable and can undermine robustness at multiple stages of the AS pipeline.

However, AS face a much broader threat landscape. **Physical attacks**–such as adversarial stickers (Eykholt et al. 2018), patches (Brown et al. 2018), or crafted objects (Kong et al. 2020)–exploit the perception pipeline by manipulating the environment itself, often defeating digital-only defenses and persisting across sensors, agents, and time.

Cross-modal and systemic attacks further challenge AS, targeting their reliance on multiple, distributed sensors and communication channels. Examples include GPS spoofing (Horton and Ranganathan 2018), LiDAR jamming (Cao et al. 2019), CAN bus manipulation (Kang et al. 2021), and attacks on federated learning (Yang et al. 2021), each capable of inducing both local and system-wide failures.

Cascading and life-cycle-aware threats are particularly critical. A single successful attack at perception can propagate via sensor fusion, scenario prediction, and control feedback loops, leading to mission-level safety breaches (e.g., semantic DoS Wan et al. 2022, adversarial planning Edelkamp 2023). These systemic vulnerabilities are largely overlooked in standard ML taxonomies.

Limitations of canonical taxonomies: Most classical frameworks categorize attacks by knowledge and timing, but largely omit the location layer, specially physical attacks and system-level propagation, reflecting a historical focus on static image classifiers and digital benchmarks. In AS, this omission is critical: physical and cross-modal threats are often the most dangerous, propagating through the stack and undermining safety in ways digital-only frameworks cannot capture. This is further pictured in appendix A, Table 10.

These limitations motivate our evaluation of attacks by location (physical and digital) developed in Sect. 3, and our life-cycle and stack-aware matrices developed in Sect. 4, which explicitly integrate both digital and physical threats at each layer and throughout the operational life-cycle of AS.

2.3 Defense mechanisms & autonomous systems

Adversarial defense research in AS has evolved rapidly, spanning mechanisms adapted from generic computer vision and those developed specifically for the unique constraints of AS. Defenses are most often categorized as proactive (e.g., adversarial training, regularization, input Pre-Processing, certification), reactive (e.g., detection, denoising, reconstruction), or, as as new category found in this review, unified approaches that integrate multiple strategies and account for the layered nature of AS deployments.

Proactive defenses such as adversarial training (Madry et al. 2019) remain foundational, retraining models on adversarial examples to improve robustness. This method, applied to both image and LiDAR-based perception modules (e.g., Lu and Radha (2023) for scaling attacks in KITTI/Waymo scenarios), demonstrates gains under known digital threats. However, these approaches incur high computational cost and generalize poorly to unseen or physical attacks, which often bypass digital adversarial defenses (Rozsa et al. 2016; Chen and Lee 2021). Additional proactive methods, including regularization (Szegedy et al. 2013; Ross and Doshi-Velez 2018), model distillation (Hinton et al. 2015; Papernot et al. 2016c), and input Pre-Processing (denoising, smoothing) (Xie et al. 2017a; Liao et al. 2018) offer marginal improvements, but often at the cost of clean accuracy or robustness to adaptive adversaries (Li et al. 2024a; Lou et al. 2023).

Model ensembles (Tramèr et al. 2017; Xie et al. 2017b) have also been explored to increase diversity and resilience, but their increased inference latency and hardware requirements are problematic for real-time AS tasks, limiting on-vehicle deployment (Lu et al. 2023; Zhao et al. 2024b). Certified defenses, including randomized smoothing (Cohen et al. 2019; Zhang et al. 2022c) and formal verification (Gowal et al. 2018; Lecuyer et al. 2019), offer provable guarantees under certain conditions, yet typically remain restricted to limited model classes and do not extend easily to full-stack or dynamic AS environments.

Reactive defenses monitor and respond to attacks at runtime. Detection-based mechanisms, such as those in Among Us Li et al. (2023) (cooperative AVs) or PhySense (Yu et al. 2024) (physical perturbation detection) use input monitoring or auxiliary detectors to identify adversarial events. While valuable, such approaches can suffer from high false positive rates and are vulnerable to sophisticated, adaptive attacks (Soares et al. 2022; Abdu-Aguye et al. 2020). Denoising and reconstruction via autoencoders or similar tools (Meng and Chen 2017; Samangouei et al. 2018) can restore clean inputs, but may introduce harmful delay or information loss–unacceptable in safety-critical AS.

Unified and stack-aware defenses are gaining attention as the limitations of layer or mechanism-specific solutions become clear. For instance, UMoE Fusion (Lou et al. 2023) exploits multimodal sensor fusion to mitigate sensor blinding, while SpecGuard (Dash et al. 2024) provides sensor and layer-aware detection against UAV sensor spoofing addressing vulnerabilities beyond the perception layer. PatchCleanser (Xiang et al. 2022) and Segment-and-Complete (Liu et al. 2022) combine certified smoothing with detection to target physical patch attacks. Temporal defenses such as ADAV (Mu 2024) and Time-Travel Defense (Etim and Szefer 2024) incorporate cross-frame and historical consistency, crucial for detecting persistent or stealthy threats in dynamic settings.

Unified defense frameworks, e.g., UniCAD (Pellicer et al. 2024), MixDefense (Du et al. 2018), and UNMASK (Freitas et al. 2020), integrate detection, denoising, and robust classification to provide scalable, adaptive defense pipelines more suitable for realistic AS opera-

tion. However, most existing defenses, even those tailored for AS, are evaluated primarily at the perception layer and fail to systematically assess downstream effects on planning, control, or mission-level safety.

The entire taxonomy and surveyed papers can be found in Appendix A, Table 11

2.4 Datasets and benchmarks for AS robustness

Effective evaluation of adversarial robustness in AS relies on benchmarks that capture both the technical complexity and real-world context in which these systems operate. The evolution of benchmarks in this space has both propelled adversarial machine learning and introduced critical challenges unique to AS contexts. Early breakthroughs in adversarial attacks and defenses were closely tied to canonical datasets such as MNIST (Lecun et al. 1998), CIFAR-10/100 (Krizhevsky 2009), and ImageNet (Deng et al. 2009). These simple, accessible, and widespread benchmarks enabled the rapid development of fundamental attack algorithms like FGSM and PGD (Goodfellow et al. 2014; Madry et al. 2019), and laid the foundation for robustness research, including systematic evaluations on corrupted or perturbed variants such as ImageNet-P (Hendrycks et al. 2021), CIFAR-C, and CIFAR-P (Hendrycks and Dietterich 2019).

Despite their foundational role, these datasets are now recognized as insufficient proxies for AS robustness due to their static, digital nature and lack of feedback, temporal dependencies, or sensor diversity. Hendrycks et al. (2021) and Croce et al. (2020) demonstrate that robustness metrics obtained on the traditional benchmarks often overstate real-world safety. Models robust on CIFAR or ImageNet may fail when confronted with the complexities of multi-modal perception, sensor fusion, or dynamic interactions in actual AS deployments. This disconnect is further underscored by simulation-to-reality transfer failures, as documented in Nesti et al. (2022); Xu et al. (2022).

To address these limitations, the field has gradually shifted towards more application-driven and AS-oriented datasets. DOTA (Xia et al. 2018) introduced complex aerial scenes and diverse object viewpoints, directly benefiting research in UAV and aerial surveillance. The Mapillary Traffic Sign Dataset (Poggi and Mattoccia 2017) captures traffic sign variation in real-world conditions, serving as a testbed for perception modules in autonomous driving. Such datasets improve environmental fidelity and task relevance but still fall short of providing holistic benchmarks for closed-loop or stack-wide robustness.

Recent advances in simulation environments–such as CARLA-GeAR (Nesti et al. 2022), SafeBench (Xu et al. 2022), and RobustE2E (Jiang et al. 2024)–have enabled holistic, closed-loop evaluation of adversarial threats across the full AS stack. These platforms support the generation of physically realizable attacks (e.g., adversarial patches, sensor spoofing), multi-agent and V2X scenarios (Li et al. 2023; Zhao et al. 2024b), and robust testing under diverse conditions (Lou et al. 2023; Zhang et al. 2023). Real-world datasets–such as Car Hacking (Kang et al. 2021) and adversarial Google Street View (Etim and Szefer 2024)–offer authentic sensor and actuator traces, though they lack the diversity and control of simulated environments.

Despite this, much adversarial research remains focused on standard vision models, with attacks like C&W (Carlini and Wagner 2017b), AutoAttack (Croce and Hein 2020), and patch-based methods (Brown et al. 2018), and defenses such as randomized smoothing (Cohen et al. 2019), MixDefense (Du et al. 2018), and certified patch segmen-

tation (Zhang et al. 2022c), almost exclusively evaluated on datasets like ImageNet or RobustBench (Croce et al. 2020). This leaves a gap in addressing how adversarial effects propagate across perception, planning, and control in realistic AS settings.

AS-specific research is bridging this divide by introducing attacks targeting the full system stack–e.g., physical patching (Eykholt et al. 2018; Li et al. 2022), LiDAR spoofing (Cao et al. 2019), sensor-fusion breakdowns (Lou et al. 2023; Zhao et al. 2024b), and CAN-bus injection (Khan et al. 2022)–and by leveraging advanced benchmarks and simulation platforms. Concurrently, new defenses emphasize multimodal anomaly detection (Lou et al. 2023), certified segmentation (Zhang et al. 2022c), physical input filtering (Lu and Radha 2023), and robust V2X fusion (Zhao et al. 2024b), increasingly targeting end-to-end, stack-aware robustness (Jiang et al. 2024).

A summarized illustration can be found in Appendix A, Table 12.

While this move toward AS-specific realism has enhanced operational relevance, it also fragments the field. Different works use incompatible sensor suites, attack models, scenario generators, and evaluation protocols—as highlighted in recent benchmark studies (Xu et al. 2022; Nesti et al. 2022). Even subtle differences in simulation parameters or the spatial/temporal configuration of physical attacks can yield markedly divergent robustness evaluations, severely limiting reproducibility and comparability across the literature. Consequently, there is a growing consensus, reflected in recent works (Croce et al. 2020; Xu et al. 2022; Lou et al. 2023). That progress depends on unified frameworks and holistic benchmarks: those that can relate algorithmic advances in general adversarial robustness to deployment in AS, and, reciprocally, that enable AS-specific innovations to be evaluated in the context of broader vision robustness objectives.

This persistent fragmentation across datasets, evaluation protocols, and adversarial methodology underscores the need for a unified approach–one that systematically bridges the gap between general computer vision research and the operational requirements of AS. To address this, our review introduces a threat-matrix-driven evaluation strategy (see Sect. 4). The unification is finally brought to fruition in in our Critical Appraisal of Defenses in the Context of Autonomous Systems, (see Sect. 5).

3 Adversarial attacks in AS: digital and physical locations

Adversarial attacks in Autonomous Systems can be broadly categorized on the basis of their location into two primary domains: digital and physical. Digital attacks occur within the digital pipeline, targeting input data or communications, while physical attacks exploit real-world environments to manipulate sensory input.

3.1 Digital attacks

Digital adversarial attacks focus on manipulating input data directly in the digital domain to deceive machine learning (ML) models. These attacks are some of the most extensively studied due to their accessibility and the relative simplicity of generating adversarial perturbations. Common methods include the aforementioned FGSM (Goodfellow et al. 2014), PGD (Madry et al. 2019), DDN (Rony et al. 2019), or Carlini and Wagner (2017b) amongst others. Figure 2 illustrates an example of FGSM.

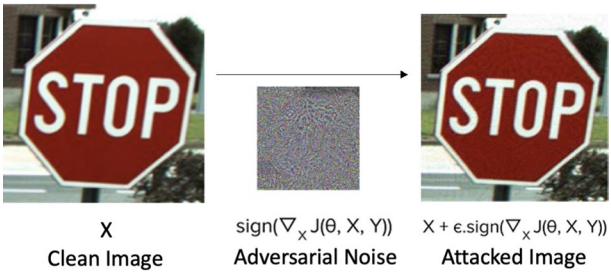

Fig. 2 Example of digital adversarial attack (FGSM)

These attacks differ in optimization strategies (e.g., single-step vs. iterative), misclassification objectives (targeted vs. untargeted), and their perturbation budget (constrained by ℓ_p norms or pixel count). Their applications in AS include not only direct evasion of perception pipelines but also poisoning training datasets, injecting malicious patterns into communication logs, or crafting precursors to physical-world attacks through digital-to-physical transfer.

Despite their digital nature, these attacks pose concrete threats to deployed systems, especially when deployed over OTA updates, V2X communication, or shared ML pipelines. As such, a clear comparative understanding of their effectiveness, stealth, and robustness is vital for evaluating the threat landscape faced by real-world autonomous platforms.

To this end, Table 3 presents a structured quantitative synthesis of the fundamental digital adversarial attacks applicable to AS-related vision models. It summarizes their success rates, perturbation magnitudes, transferability across models, and contextual relevance.

3.2 Physical attacks

Physical adversarial attacks are a type of attack in which an adversary attempts to deceive or mislead a ML approach that relies on data gathered from the environment through the use of physical hardware sensors such as cameras. Physical attacks do so by introducing physical perturbations to its environment or inputs. Physical adversarial attacks can take various forms, such as altering the lighting conditions (Xiao et al. 2018), modifying the appearance of objects in the environment (Oslund et al. 2022), or manipulating the sensors that the autonomous system relies on to perceive the world (Cao et al. 2019). Furthermore, in many cases, attacks may be unnoticeable to humans when placed in the real world as they may be mistaken by decorations, urban art or vandalism and not seen as a bigger threat by humans, which hinders the possibility of manual human intervention to prevent attacks in real time. Physical attacks can be configured both in a white-Box or a Black-box setting with differences in performance based on the attack, and their timing would normally be considered Evasion, although it could be the case that they act as Poisoning attacks in the event that the system being compromised is in the learning stage.

Physical adversarial attacks can be generated by transferring digital adversarial attacks into physical objects as demonstrated in various studies (Kurakin et al. 2016; Athalye et al. 2017; Sharif et al. 2016). Different techniques to achieve that shift exist which obtain different levels of attack robustness. However, in the physical environment, attack robustness is challenged by other factors, including natural changes in environment conditions, the attack surface being smaller and more complex due to it being three dimensional, the background not being alterable, or different camera angles.

Table 3 Quantitative overview of digital adversarial attacks targeting autonomous systems

Attack method	Description	Common AS targets	Success rate (%)	Perturbation size	Robustness (transferability)	Remarks
FGSM (Goodfellow et al. 2014)	Fast one-step gradient sign method. Efficient but weaker	CNNs in perception pipelines (YOLO, ResNet, MobileNet)	65–85%	$\ell_\infty \leq 0.03$	Low (model-specific)	Computationally fast, non-iterative
I-FGSM / PGD (Kurakin et al. 2016; Madry et al. 2019)	Iterative FGSM; PGD is a universal first-order attack	Traffic sign classifiers, camera input stream	90–99%	$\ell_\infty \leq 0.03$	Medium (higher with ensemble)	Standard benchmark for robust training
DDN (Rony et al. 2019)	Minimizes norm directly via decoupled optimization	Perception tasks (ResNet, EfficientNet)	80–95%	$\ell_2 \approx 0.5$	Medium	Good for precise attack with minimal distortion
C&W (Carlini and Wagner 2017b)	Optimizes distortion with a Lagrangian framework. Very strong	Sensor fusion, camera input, LiDAR projection classifiers	95–100%	$\ell_2 \approx 0.1$ or lower	High	Slow but stealthy; often bypasses defenses
DeepFool (Moosavi-Dezfooli et al. 2016)	Minimal ℓ_2 perturbation to cross decision boundary	AS camera classifiers, edge detectors	85–95%	$\ell_2 \approx 0.01$–0.1	Medium	Produces very imperceptible noise
UAP (Moosavi-Dezfooli et al. 2017)	Image-agnostic perturbations that generalize across inputs	Scene classification (e.g., road conditions)	80–92%	$\ell_2 \leq 0.3$	High	Transferable to unseen data and models
JSMA (Papernot et al. 2016)	Perturbs salient pixels using gradient-based saliency maps	AS object detectors	70–90%	Few pixels ($< 1\%$)	Low	High distortion when success is enforced
Square Attack (Andriushchenko et al. 2020)	Score-based black-box attack with local square updates	On-device perception models	85–95%	$\ell_\infty \leq 0.03$	Medium	Efficient in query-limited settings
SimBA (Guo et al. 2019b)	Black-box attack via randomized low-frequency noise directions	Control layer feature extractors	75–90%	$\ell_2 \leq 0.5$	Medium	Simple and effective in low-query regime
One-Pixel / Few-Pixel (Su et al. 2019; Xiao et al. 2018)	Changes only one or few pixels. Evasion with minimal footprint	Simple classifiers (MNIST, GTSRB)	30–70%	1–5 pixels	Very Low	Not robust; poor scalability to complex images

Table 3 (continued)

Attack method	Description	Common AS targets	Success rate (%)	Perturbation size	Robustness (transferability)	Remarks
Backdoor (e.g., BadNets) (Gu et al. 2017)	Inserts triggers into training data. Attack triggered only when pattern appears	Entire AS training pipelines	100% (when triggered)	Trigger patch (0.5–5% area)	High (persistent)	Remains dormant; extremely dangerous in safety-critical AS
MetaPoison (Huang et al. 2020)	Craft poisoned training data to manipulate decision boundaries	Offline AS model training (perception)	80–95%	Clean-label (stealth)	High	Invisible to defenders; long-term threat

Success rate (%) reflects attack effectiveness reported across standard AS-relevant models and datasets. *Perturbation size* describes typical norm-bound constraints (e.g., ℓ_∞, ℓ_2) or pixel counts. *Robustness* refers to transferability across models, datasets, and tasks. Metrics are extracted or averaged from controlled benchmarks and attack papers, focusing on vision-based perception pipelines in AS

In the context of AS, physical adversarial attacks represent a significant hazard, with the potential to compromise system safety and dependability. For instance, autonomous vehicles could be misled into misinterpreting traffic control devices such as stop signs or traffic lights, precipitating a potentially perilous situation. The effectiveness of physical adversarial attacks on object detection systems, pivotal in autonomous vehicles, was demonstrated in a study by Eykholt et al. (2018). The research indicated that a physical evasion attack could be orchestrated by adding minimal perturbations to stop signs, thereby distorting the accurate perception of autonomous vehicles.

There are diverging views within the community regarding the effectiveness of these physical adversarial perturbations. Some studies, such as Lu et al. (2017), suggest that while these adversarial alterations could lead a deep neural network to misinterpret a stop sign image in a physical environment within a specific range of distances and angles, they are not uniformly successful in duping object detectors across varied distances and viewing angles. However, it should be noted that these experiments were conducted in a simplified setting, involving printed attack signs.

More sophisticated and resilient attack methods have since emerged, capable of handling changes in viewpoint, some of which are further explored in this paper. Moreover, it is suggested that as AS and the various deep learning methodologies underpinning their operation continue to evolve, the nature of attacks will similarly adapt and become more advanced. Therefore, contrary to some researchers who may downplay the potential harm of physical adversarial attacks, these threats are considered critical and warrant urgent attention in order to ensure system integrity and safety. A summary of the main types of physical attacks is displayed at the end of this section in 4.

3.2.1 Adversarial stickers and paintings

The use of adversarial stickers and paintings for deceiving object detection or image classification in AS has been a topic of study. Specifically, Eykholt et al. (2018) examined their effectiveness on deep learning models used in autonomous vehicles. The method involves placing carefully crafted stickers for target objects into the real world, which can cause

Table 4 Comprehensive summary of physical adversarial attacks applicable to autonomous systems, integrating both qualitative and quantitative evidence from the literature

Attack type	Target system(s)	Description	Implications	Success rate (%)	Size	Robustness	Key studies
Adversarial Stickers	Traffic Sign Recognition, Object Detection	Printed perturbations (e.g., on signs) crafted using FGSM or GANs to mislead perception models	Misclassification of traffic signs; risk of safety-critical errors in AVs	Up to 91.49%	$\leq$5% of object area	Partial (angle/distance sensitive)	Eykholt et al. (2018); Oslund et al. (2022); Zhu et al. (2024)
Adversarial Patches (2D)	UAV Detection, Person Detection	Small 2D patches embedded in clothing or scenes, optimized to evade detection	Enables human evasion from surveillance or drone systems	Up to 90%	$\leq$1% of image area	Limited	Thys et al. (2019); Wu et al. (2020)
Adversarial Patches (3D)	Object Detection (YOLO, SSD)	Physically printed 3D patches placed on real objects (e.g., vehicles)	Persistent misclassification of camouflaged objects	Up to 85%	Object surface dependent	High (real-world tested)	Toheed et al. (2022); Du et al. (2022)
Adversarial Objects (2D)	Image Classification	Printed adversarial images misclassified under varied conditions	Demonstrates real-world vulnerability of classifiers	65–85%	Full object	Partial	Kurakin et al. (2016)
Adversarial Objects (3D)	Object Detection, Multi-Sensor Fusion Systems	Crafted 3D shapes optimized via EOT or end-to-end sensor-aware learning	Compromises multi-sensor fusion in autonomous vehicles	80–90%	Full object	High	Athalye et al. (2017); Cao et al. (2020)
Adversarial Billboards	Autonomous Driving Systems	Adversarial large-scale signs created via optimization in 3D simulator	Attacks AS from afar; misguides perception in motion	Approximately 65% misdetection	Full billboard	Medium	Zhou et al. (2020)

Table 4 (continued)

Attack type	Target system(s)	Description	Implications	Success rate (%)	Size	Robustness	Key studies
Adversarial Clothing	Person Recognition	T-shirts or jackets with adversarial patterns to evade detection	Enables physical anonymity from AI-based surveillance	57–74%	Clothing-scale	Partial	Wu et al. (2020)
Adversarial Rain	Object Detection, Classification	Raindrop overlays on lens or images to obstruct vision systems	Misinterpretation of surroundings under weather conditions	60–70% accuracy drop	N/A	Medium	Guesmi et al. (2023)
Adversarial Lighting	Object Detection	Controlled lighting (e.g., glare/shadow) to cause detection failures	Disrupts feature extraction; breaks perception	Up to 93.7% fooling rate	Global	High (controlled)	Hsiao et al. (2024)

Target system(s) refers to the machine learning subsystems being attacked (e.g., traffic sign recognizer, object detector). *Success rate (%)* indicates the reported attack success under physical-world or simulation conditions. *Size* estimates the spatial footprint of the adversarial pattern relative to the object or image surface. *Robustness* denotes the resilience of the attack to changes in viewpoint, lighting, and physical conditions. Metrics are synthesized from experimental results in the cited studies; where multiple results are reported, the maximum or typical observed value is given

Fig. 3 Example of adversarial stickers

misclassification of the object detection system. The authors demonstrated that these stickers could be designed to be virtually imperceptible to humans, but still deceive the object detection system. A visualization of the attack is shown in Fig. 3

To generate the adversarial stickers and paintings, the authors used a modified version of the FSM algorithm. They began by selecting a target label, such as a yield sign or a speed limit sign, and used the FGSM algorithm to generate a small perturbation that would cause the object detection system to misclassify the stop sign as the target label. The authors also used a generative adversarial network (GAN) to train a model that could generate images that looked similar to stop signs but contained the adversarial perturbations, while remaining imperceptible to humans.

The study's findings suggest that the adversarial stickers succeeded in deceiving numerous cutting-edge deep learning models employed in autonomous vehicles, resulting in potentially perilous circumstances. Importantly, the researchers demonstrated the transfer-

ability of these adversarial stickers across disparate models and camera types. Furthermore, the study investigated the influence of physical factors such as lighting conditions, viewing angles, and distances, on the effectiveness of the adversarial stickers. The effectiveness of the stickers did exhibit variation depending on these factors, but crucially retained effectiveness across a broad spectrum of scenarios.

3.2.2 Adversarial patches

Adversarial patches refer to intricately crafted patches that can be introduced into an image to misguide object detection systems and cause them to misclassify objects in the scene. Such attacks have been previously used to prevent cameras from detecting humans, as evidenced by the development of T-shirts that are printed with adversarial patches (Wu et al. 2020) or by having people wear the patches themselves (Thys et al. 2019). In addition to this, adversarial patches have also been utilized to evade face recognition systems (Komkov and Petiushko 2021) or to prevent AS from detecting objects in the scene (Du et al. 2022).

Work by Zhang et al. (2022a) explores the vulnerability of multi-scale object detection models utilized in UAVs to adversarial patch attacks. The authors, similarly to the way adversarial stickers are generated, employed a modified version of the fast gradient sign method (FGSM) algorithm to generate adversarial patches. They initially trained a deep learning model to create patches that could be incorporated into an image to induce misclassification by the object detection system. The patches were designed to be small and inconspicuous to humans but yet potent in deceiving the object detection system.

The research found that adversarial patches were efficient in deceiving several cutting-edge object detection models employed in UAVs. The authors showed that even when the patches covered less than one percent of the image area, they could still deceive the object detection system. Furthermore, the patches were transferable across different object detection models, making them a potential threat to UAVs that rely on deep learning models for object detection.

The research also scrutinized the impact of the size and location of the adversarial patches on the attack's effectiveness. The authors found that larger patches and patches placed in more critical areas of the image were more effective in deceiving the object detection system.

It is worth noting that a potential limitation of the study at hand is that the patch experiment results only demonstrate the path being 2D and placed on top of the image. However, in real-world scenarios, attackers are more likely to use these patches to camouflage objects, such as military vehicles like tanks or fighter jets with an adversarial patch. Therefore, the use of a 3D adversarial patch may be more realistic in such situations.

To address this limitation, Toheed et al. (2022) proposes a method for conducting physical adversarial attacks on object detection systems using 3D adversarial objects. The authors argue that current adversarial attacks on object detectors mainly rely on 2D adversarial perturbations, which have limited ability to cause misclassification of objects in the real world.

The authors introduce a 3D adversarial object that is designed to be imperceptible to humans but can cause misclassification of objects by the object detector. The 3D object is created using computer-aided design (CAD) software and 3D printing technology. The proposed attack is tested on the YOLOv2 object detection system and the COCO dataset, demonstrating its effectiveness in causing misclassification of objects in the real world.

3.2.3 Adversarial objects

Adversarial objects are crafted in a way that they cause the ML model to misclassify, misinterpret, or fail to recognize them, even though they might appear normal to the human eye. They follow a similar approach to adversarial stickers or patches. However, they differ in that a complete 2D or 3D object is built.

Kurakin et al. (2016) was one of the first to investigate 2D physical adversarial objects, this paper investigates the effectiveness of adversarial examples in real-world settings. The authors focus on the transferability of adversarial examples between digital and physical domains, as well as their robustness to various transformations, such as changes in camera angle and lighting conditions. The authors extend their investigation to the physical world, questioning whether adversarial examples generated in the digital domain can still be effective when captured by a camera and processed by a ML model.

To study this question, the authors generate adversarial examples using FGSM and print them out, simulating a physical-world scenario. They then capture images of these printed adversarial examples using a smartphone camera and feed the captured images to a deep learning model to evaluate the model's performance.

The experiments show that adversarial examples generated in the digital domain can still be effective in the physical world, causing the ML model to misclassify the printed images. The authors also demonstrate that the adversarial examples are robust to various transformations, such as changes in camera angle, lighting conditions, and resizing of the images. This finding suggests that adversarial examples pose a significant challenge to the deployment of deep learning models in real-world applications, as they can cause the models to make incorrect decisions even under different physical conditions.

More curated and targeted to Autonomous System papers in the 2D object landscape include (Kong et al. 2020; Zhou et al. 2020). Zhou et al. (2020) presents a systematic approach for generating adversarial billboards designed to compromise object detection models in autonomous driving systems. The authors propose a bi-level optimization framework that considers both the attack's success probability and the perturbation's perceptual similarity. They leverage a 3D simulator to account for physical-world factors such as lighting, camera perspective, and occlusion. While this approach provides valuable insights into the robustness of object detection models under various physical-world scenarios, the use of a 3D simulator may not fully capture the complexity of real-world conditions, potentially limiting the generalizability of the results. Kong et al. (2020) employs a Generative Adversarial Network (GAN) to create adversarial examples resilient to real-world environmental factors. The method comprises a generator network responsible for producing adversarial perturbations and a discriminator network tasked with discerning between real and adversarial examples. To enhance the transferability of the generated adversarial examples, the authors incorporate domain adaptation techniques and apply geometric and photometric transformations during training. While Kong et al. (2020) demonstrates the potential for crafting physical-world-resilient adversarial examples, the adversarial training process can be computationally expensive and sensitive to hyperparameters, which may limit its practical applicability.

Athalye et al. (2017) was one of the first works to introduce 3D adversarial objects. The paper presents a novel approach to generating adversarial examples that are robust to various transformations and are effective in both the digital and physical domains. The authors pro-

pose a method called Expectation over Transformation (EOT), which aims to create adversarial examples that maintain their adversarial properties under different transformations.

Traditional adversarial example generation methods often focus on fooling a ML model in the digital domain, without considering the effects of real-world transformations, such as rotations, translations, and changes in lighting. As a result, these adversarial examples may lose their effectiveness when applied to physical objects or real-world scenarios. To address this issue, the authors introduce the EOT algorithm, which incorporates an expectation over a chosen set of transformations during the adversarial example generation process. By optimizing the adversarial perturbation under this expectation, the algorithm ensures that the generated adversarial examples are robust to the specified set of transformations.

The authors evaluated the performance of the EOT algorithm on various state-of-the-art deep learning models, such as Inception v3 and ResNet, using different datasets like ImageNet and CIFAR-10. They also compare the EOT algorithm with other existing methods, such as FGSM and PGD. The results demonstrate that the EOT algorithm is able to generate adversarial examples that are robust to a wide range of transformations, outperforming other methods in both digital and physical domains. The authors further showcased the effectiveness of the EOT algorithm through real-world demonstrations, such as 3D printed objects and images displayed on a screen.

Cao et al. (2020) specifically targets the vulnerabilities of autonomous driving systems to 3D adversarial objects. This paper specifically targets Multi-Sensor Fusion (MSF)-based perception systems used in autonomous vehicles. The authors propose a real-time, end-to-end optimization algorithm that takes into account the physical constraints and sensor characteristics of the MSF-based perception system to generate 3D adversarial objects. By considering the limitations of the sensors and the physical constraints of the objects, the proposed method generates adversarial objects that can deceive the MSF-based perception system in real-world scenarios. The paper evaluates its method using simulation and real-world experiments, focusing on the effectiveness of the 3D adversarial objects in deceiving MSF-based perception systems in autonomous vehicles.

Table 4 summarizes the main types of physical adversarial attacks, their implications, and key examples along with simple quantitative indicators such as Success Rate or Robustness to further contextualize their relevance in vision models and therefore to AS.

4 Threat modelling in autonomous systems

This section presents a comprehensive framework for threat modeling in AS, with a particular focus on vision-based models. We introduce a taxonomy that systematically analyzes the exposure of each stage in the AS life-cycle to adversarial attacks (both digital and physical). By mapping specific attack vectors to corresponding life-cycle components and system layers, this framework provides a structured basis for identifying vulnerabilities and informs the development of effective, targeted defense strategies for real-world AS deployments.

4.1 Life-cycle attack matrix

We introduce the *AS AI Life-Cycle Attack Matrix* (see Fig. 4), a framework that systematically categorizes adversarial threats targeting AS across the Data, Training, and Inference

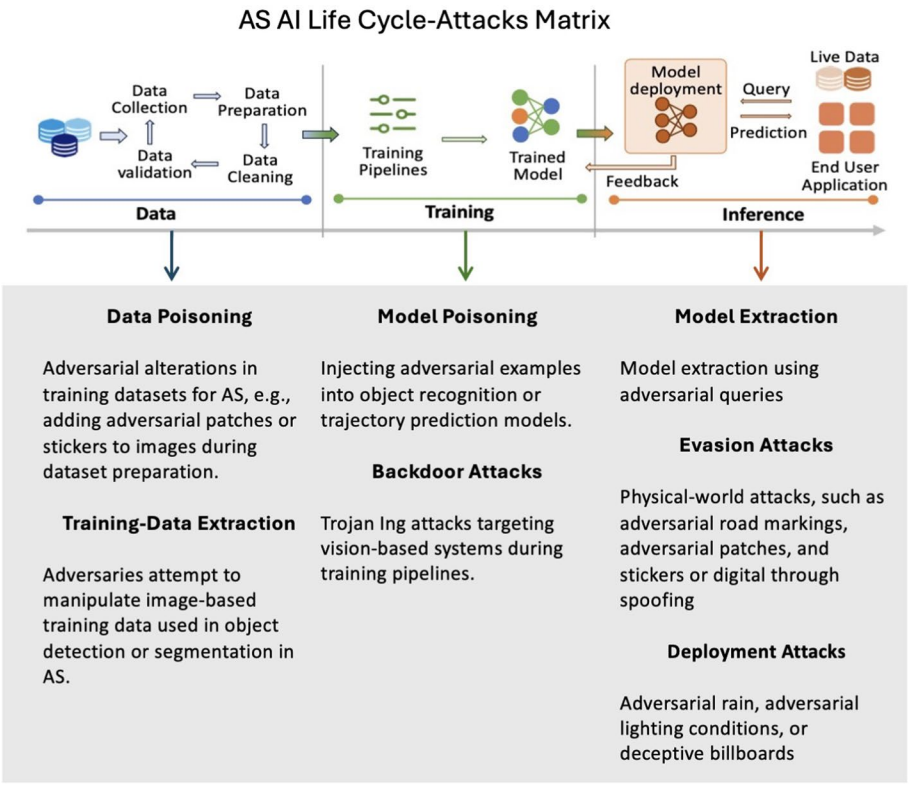

Fig. 4 AS AI life-cycle attack matrix

stages of the AI life-cycle. By mapping attack types to each stage, the matrix provides a comprehensive structure for identifying vulnerabilities, understanding how adversaries exploit phase-specific weaknesses, and informing the design of more effective defense strategies.

Figure 4 organizes adversarial threats into three main stages of the AI life-cycle: Data, Training, and Inference. Each stage is associated with characteristic attack types that leverage distinct vulnerabilities in AS pipelines.

At the **Data stage**, adversaries may engage in:

- **Data poisoning attacks:** Introducing malicious data into the training dataset to corrupt the learning process, leading to erroneous model behavior. For instance, altering traffic sign images to mislead recognition systems in autonomous vehicles (Morgulis et al. 2019).
- **Training-data extraction attacks:** Extracting sensitive information from the training data, potentially compromising privacy and security. This can involve reconstructing proprietary datasets used in AS development (Malik et al. 2024).

During the **Training stage**, potential attacks include:

- **Model poisoning attacks:** Manipulating the training process to embed vulnerabilities

within the model, which can be exploited during deployment. This includes tampering with the training data or the learning algorithm itself (Almutairi and Barnawi 2023).

- **Backdoor attacks:** Inserting hidden triggers into the model that cause it to behave maliciously when specific conditions are met. For example, embedding triggers that activate under certain visual patterns encountered by AS (Pourkeshavarz et al. 2024).
- **Attacks in federated learning (FL):** Federated learning offers a decentralized approach to training machine learning models across multiple devices, making it particularly relevant for AS applications such as autonomous vehicles. In FL, each client–such as an autonomous vehicle–trains a local model using its own data. Only the model updates are shared with a central server, where they are aggregated to create a global model. This approach not only preserves data privacy but also reduces computational and communication costs by distributing the training process across multiple devices (Jallepalli et al. 2021).
  However, FL's decentralized nature introduces unique security challenges. Malicious actors can exploit the collaborative training process to compromise the global model. For instance, a rogue client might poison its local training data or tamper with model updates, leading to degraded performance or targeted misbehavior. Moreover, FL's privacy-preserving mechanisms, such as secure aggregation and differential privacy, can make detecting such attacks more difficult, further complicating the task of ensuring robust security. Recent studies, including (Li et al. 2024c; Queyrut et al. 2023; Shi et al. 2022), provide a comprehensive overview of FL architectures, their adversarial challenges, and potential defense strategies within AS. A simple visualization of attack vectors in a FL architecture is shown in Fig. 5.

At the **Inference stage**, AS are susceptible to:

- **Model extraction attacks:** Adversaries query the deployed model to reconstruct its parameters or architecture, facilitating intellectual property theft or enabling further attacks (Malik et al. 2024).
- **Evasion attacks:** Crafting inputs that are intentionally designed to be misclassified by the model, thereby bypassing security measures. Physical-world examples include adversarial patches or stickers that cause misclassification in object detection systems (Girdhar et al. 2023).
- **Prompt attacks:** Exploiting prompt-based systems by injecting malicious prompts that

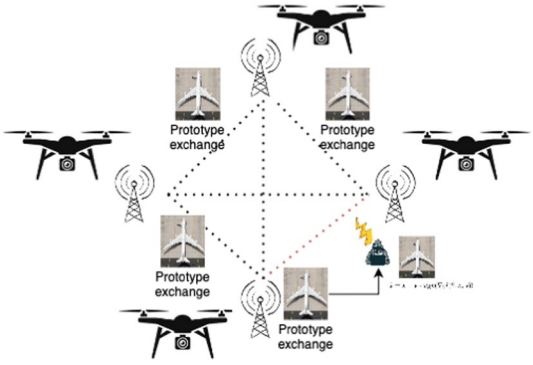

Fig. 5 Example of prototype-based FL architecture and attack surface

alter the model's behavior or outputs, potentially leading to unintended actions in AS (Shan et al. 2024).

- **Adversarial deployment attacks:** Introducing adversarial elements into the environment, such as deceptive road markings or manipulated traffic signs, to mislead the AS perception and decision-making processes (Boltachev 2024).

This taxonomy underscores the multifaceted nature of adversarial threats across the AI life-cycle in Autonomous Systems. By systematically categorizing these attacks, we aim to enhance the understanding and development of robust defense mechanisms tailored to each stage of the AI deployment pipeline.

4.2 Exposure–impact matrix

The *AS Adversarial Exposure-Impact Matrix*, illustrated in Fig. 6, offers a detailed taxonomy of adversarial attack vectors that specifically exploit vulnerabilities in AS. The matrix organizes these vulnerabilities according to fundamental AI challenges, such as the need for large datasets, sensitivity to model updates, similarities across models, and input fragility, linking each to concrete attack surfaces, including data pipelines, model APIs, and environmental inputs.

These vulnerabilities enable a wide spectrum of attacks, ranging from *data poisoning* and *backdoors* during training to *model extraction* and *evasion* at inference. The matrix clarifies both where and how AS can be compromised and traces the downstream consequences from data collection and model preparation through deployment to operational harms such as misguidance, sabotage, or intellectual property (IP) theft.

AI inherent vulnerabilities and attack surfaces: AS inherit several critical vulnerabilities from the underlying AI models and datasets on which they rely, exposing multiple attack surfaces:

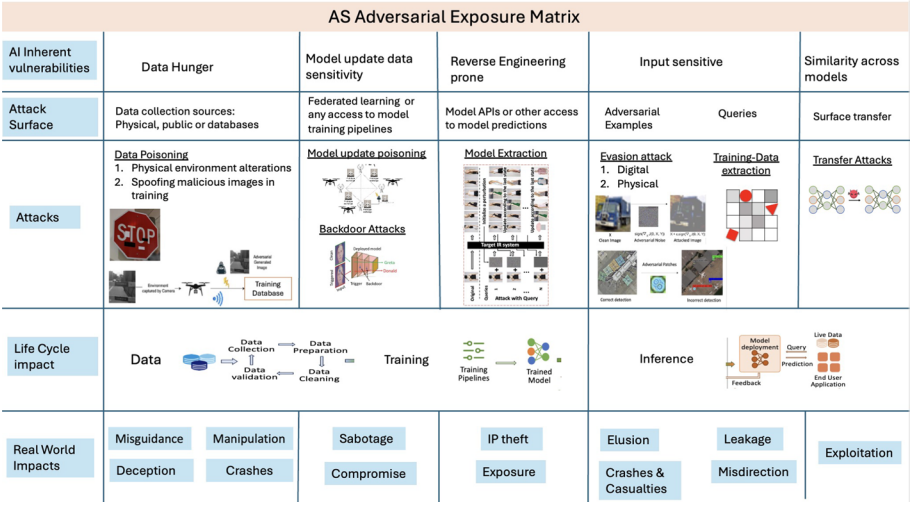

Fig. 6 AS exposure-impact matrix

- **Data hunger:** The requirement for large and diverse datasets makes AS vulnerable to *data poisoning attacks*, where adversarial modifications–such as altered traffic sign images–are injected into the training data (Eykholt et al. 2018).
- **Model update sensitivity:** The adoption of federated learning and access to model update pipelines introduce the risk of *model poisoning attacks*, allowing adversaries to manipulate updates and embed *backdoors* (Cheng et al. 2021).
- **Input sensitivity:** The inherent fragility of AI models to subtle input changes makes them susceptible to *adversarial examples*, including both digital perturbations and physical attacks (such as stickers or patches on objects) (Brown et al. 2018).
- **Similarity across models:** The resemblance between different models allows for *transfer attacks*, where adversarial examples crafted for one model can successfully mislead another (Tramèr et al. 2017).

Real-world impacts: The AS Adversarial Exposure Matrix reveals how the convergence of AI vulnerabilities, attack surfaces, and adversarial tactics results in tangible real-world consequences. By mapping these threats from data collection through training, inference, and deployment, the matrix highlights clear pathways through which Autonomous Systems can be undermined:

- **Data hunger → Data poisoning:** The demand for extensive, diverse datasets exposes AS to *data poisoning*, where physical or digital manipulation of training data causes *misguidance* and *deception* at the perception layer.
- **Model update sensitivity → Model poisoning and backdoor attacks:** Continuous model refinement in AS creates opportunities for adversaries to introduce *model poisoning* or embed *backdoors* via tainted updates. This results in *manipulation* and *sabotage*, eroding model integrity and reliability.
- **Reverse engineering prone → Model extraction:** When attackers gain access to model outputs through open APIs or similar interfaces, they can perform *model extraction*, leading to *IP theft* and *exposure* of proprietary algorithms. This undermines competitive advantage and may facilitate further adversarial actions.
- **Input sensitivity → Evasion attacks and training-data extraction:** Systems that rely on accurate sensor interpretation or user input are vulnerable to *evasion attacks* and adversarial queries. Such *elusion* and environmental *manipulation* can cause *crashes*, *casualties*, and information *leakage*, as the AS fails to interpret its environment correctly.
- **Similarity across models → Transfer attacks:** Exploiting similarities among models, adversaries can launch *transfer attacks* that scale across multiple AS platforms, resulting in widespread *exploitation* and a further erosion of public trust in these technologies.

By mapping each vulnerability and attack type to its downstream impact, the matrix underscores that even subtle technical manipulations can cascade into severe, real-world consequences. Understanding these relationships is crucial for designing robust defense strategies that ensure the reliability, safety, and integrity of Autonomous Systems.

Table 5 consolidates recent research that exemplifies the real-world impacts identified in the AS Adversarial Exposure Matrix. These studies provide concrete evidence of adversarial attacks, their methodologies, and their consequences for AS, emphasizing the need for comprehensive defense mechanisms.

Table 5 Representative set of attacks and their real-world impacts in Autonomous Systems

Study	Attack type	Real-world impact
Dynamic Adversarial Attacks on Autonomous Driving Systems Chahe et al. (2023)	Physical adversarial patches on moving objects	Misclassification of traffic signs, leading to misguidance and deception
Adversary ML Resilience in Autonomous Driving Through Human-Centered Perception Mechanisms Shah (2023)	Physical attacks on road signs (e.g., tape, graffiti)	Misclassification, causing safety hazards
Embodied Adversarial Attack: A Dynamic Robust Physical Attack in Autonomous Driving Wang et al. (2023b)	Laser-based dynamic physical attacks	Misinterpretation of the environment, resulting in potential crashes
Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems Wang et al. (2023c)	Transfer attacks leveraging model similarities	Scaled exploitation across multiple autonomous systems
Towards Robust and Secure Embodied AI: A Survey on Vulnerabilities and Attacks Xing et al. (2025)	Adversarial manipulation of AI-controlled robots	Safety-critical failures, including crashes and casualties
Discovering Adversarial Driving Maneuvers Against Autonomous Vehicles Song et al. (2023)	Adversarial driving maneuvers	System misguidance, crashes, and operational compromise
Efficient Adversarial Attack Strategy Against 3D Object Detection in Autonomous Driving Chen et al. (2024b)	3D object detection manipulation	Misclassification of objects, leading to potential crashes
Adversarial Backdoor Attack on Trajectory Prediction Pourkeshavarz et al. (2024)	Clean-label data poisoning	Causes systematic errors in path prediction, increasing collision risks

4.3 Stack–threat matrix

Because AS operate in uncontrolled, open environments, they are especially vulnerable to attacks that target the physical world. *Physical adversarial attacks* are particularly critical, as they directly compromise the perception capabilities of sensors and cameras, thereby undermining all subsequent layers. Nonetheless, vulnerabilities are not limited to physical inputs. Table 6 provides our matrix mapping relevant examples with their scenarios and implications per stack layer. Some more in depth conceptual examples are presented bellow to further understand the relevance per layer:

At the **the Perception Layer**, attacks can manipulate the sensory input of an AS, causing the system to perceive incorrect or misleading information. Adversarial attacks in computer vision can cause an AS to misclassify objects in the environment, leading to incorrect or unsafe actions (Ai et al. 2021; Wang et al. 2021).

Tampering with the perception layer often involves that further layers (planning and control) will also be compromised as data flows from one layer to the other, an incorrect view

Table 6 Stack-threat matrix

Attack description	Scenario	Target	Implications	References
Perception layer				
Adversarial Patch	Patch embedded on road sign or object	Object detector/classifier	Misclassification, system malfunction, hazardous incidents	Brown et al. (2018)
Adversarial Sticker	Adversarial sticker on object/surface	Object detection/segmentation	Scene misperception, incorrect decisions	Chen et al. (2019)
Adversarial Apparel	Person wearing adversarial clothes or accessories	Human/object classification	Pedestrian missed, security breach	Xu et al. (2020; Sharif et al. (2019)
Adversarial Object	Placing adversarially-engineered 2D/3D object in environment	Object recognition	False identification, safety risks	Kong et al. (2020)
Lighting Attack	Manipulate scene lighting/shadows	Vision-based perception	Misclassification, detection failure	Hsiao et al. (2024)
Adversarial Rain	Raindrop patterns on lens/image	Vision-based perception	Degraded perception, environmental misinterpretation	Guesmi et al. (2023)
Adversarial Clothing	Clothing designed to fool detector	Person detection/recognition	Security risk, evasion of detection	Hu et al. (2023a)
Remote Perception Attack	Malicious pattern injection via compromised comms	Camera-based detection	False negatives for critical objects	Man et al. (2023)
LiDAR Spoofing	Fake laser signals to LiDAR sensor	LiDAR perception	False obstacle detection, collision risk	Cao et al. (2019)
Planning layer				
Traffic Sign Attack	Subverted or altered traffic sign	Traffic sign recognition	Misnavigation, rule violation, accident risk	Eykholt et al. (2018)
GPS Spoofing	Falsified GPS signals	Navigation system	Route deviation, loss of control, accidents	Horton and Ranganathan (2018)
UAV Tracking Attack	Compromised tracking data or communication	UAV route/target tracker	Loss of target, mission failure	Fu et al. (2022)
Adversarial Billboard	Adversarial billboard/sign in environment	Object detection/classification	Scene confusion, misbehavior, planning error	Zhou et al. (2020)
Adversarial Planning	Crafted planner input/feedback	Planning algorithm	Unsafe/inefficient routing, increased risk	Edelkamp (2023)
Trajectory Attack	Adversarial input to prediction model	Trajectory prediction	Wrong agent movement forecast, collision	Cao et al. (2022)

Table 6 (continued)

Attack description	Scenario	Target	Implications	References
Semantic DoS Attack	Benign object induces overly conservative behavior	Behavioral planning module	Unnecessary stops or detours, degraded performance	Wan et al. (2022)
Control layer				
UAV OD Spoofing	Spoofed images for UAV detection	UAV object detection/control	Erroneous control action, unsafe maneuvers	Tian et al. (2022)
Semantic Exploit	Malicious image for segmentation/detection	Control subsystem	Poor control decisions, potential accidents	Xie et al. (2017b)
Trojaning Attack	Injecting backdoor during model training	Control algorithm	Unauthorized actuation, hijack risk	Cheng et al. (2021)
Model Extraction	Query-based model stealing	Control algorithm/model	IP theft, enables further attack planning	Li et al. (2021)
Flying Patch	Drone delivers adversarial patch into field of view	Vision-based control	Remote error injection, loss of control	Hanfeld et al. (2023)
GhostImage Attack	Remote projection of adversarial pattern	Camera-based control	Misclassification, control errors	Man et al. (2020)
CAN Injection	Malicious CAN bus message injection	Vehicle control systems	Unauthorized control, theft	Khan et al. (2022)

of the environment can lead to, for instance an incorrect route being planned and wrong commands sent to the actuators in the control layer. The scenarios for attacks that target the perception layer involve the exploitation of the area in which camera sensors actuate, in this case the physical environment, thus the threat to be considered are physical adversarial attacks. These include adversarial patches, objects and stickers which have been outlined previously and summarized in Sect. 3.2.

Attacks in to the perception layer and to other layers can be distinguished based on the attacker objectives, this means that although every successful physical attack involves alterations to the perception of the environment produced at the perception layer, not every physical attack shall be considered a perception layer attack.

Perception layer attacks aim to remove or add elements to the system's perception of the world, altering its fundamental behavior. If the example of driverless cars is considered, attacks involving adversarial traffic signs (Morgulis et al. 2019) might be more appropriately classified as planning layer attacks rather than perception layer attacks. This is because even if a stop sign is misclassified as a 45 mph speed limit sign, the car will still be able to navigate the road and recognize that a traffic sign is present. However, its planned route or correct trajectory will be altered due to an unintended decision made at the planning layer. In contrast, an attack involving a pedestrian wearing an adversarial T-shirt (Xu et al. 2020) should be considered a perception layer attack, as it renders an element invisible, preventing the car from accounting for all elements on the road. Therefore, attackers' aiming purely at the perception layer will normally leverage physical attacks targeting object detectors.

At the **Planning Layer** adversarial attacks can be crafted leveraging vulnerabilities in Deep learning classifiers including physical attack such as adversarial traffic signs as demonstarted by Morgulis et al. (2019). The security implications can include incorrect routes, traffic violations, or accidents. In Fu et al. (2022), an adaptive adversarial attack on real-time Unmanned Aerial Vehicle (UAV) tracking systems is introduced. The authors devise the Ad2Attack method, a mechanism that produces adversarial examples aimed at deceiving deep learning-powered UAV tracking systems. A successful compromise of the tracking system's performance can lead to the UAV losing track of its intended target. This loss of tracking can, in turn, result in inaccurate or suboptimal route planning, thus posing significant operational challenges.

Other significant vulnerabilities can be found in planning systems which involve gathering information from external sources to make planning decisions, such as GPS spoofing, an example of such attack can be found in Horton and Ranganathan (2018), attacks such as this can manipulate the drone's perceived location and potentially take control of its movements. Although this example is not in the image domain, it is believed that systems may use other information in the planning layer such as saved streetview images downloaded from an external server to aid navigation. Thus, attacks similar to GPS spoofing, where malicious images are injected into the planning layer leveraging wireless technology vulnerabilities, may exist in the future.

For the **Control Layer**, Tian et al. (2022) presents an architecture for an unmanned aerial vehicle (UAV) is described, in which the drone's camera acts as a sensor and sends real time images to the controller for processing and display through a Wi-Fi network. The controller, which is based on Dronet, processes the image to gain situational awareness of the environment and generates control instructions. These instructions are transmitted to the actuator through the Wi-Fi network to control the drone. Given the vulnerabilities in Wi-Fi networks, there may exist an active attacker who controls the Wi-Fi link and generates imperceptible perturbations (adversarial examples) to images sent by the camera to remain undetected. This attack may result on the drone receiving wrong velocity commands which could make it intentionally crash to an object or even a human, or at least alter its normal course. A illustration of this attack is shown in Fig. 7.

Xie et al. (2017b) explores adversarial attacks on deep learning-based semantic segmentation and object detection systems, both of which play a critical role in the control layer of autonomous vehicles. Through the generation of adversarial examples, these systems can be manipulated, leading to erroneous control decisions with potentially hazardous outcomes such as accidents or system malfunctions.

The researchers present a method for creating adversarial examples that effectively deceive both semantic segmentation and object detection algorithms. The technical backbone of this paper involves the resolution of an optimization problem, the goal of which is to create adversarial perturbations that maximize the target model's loss function while

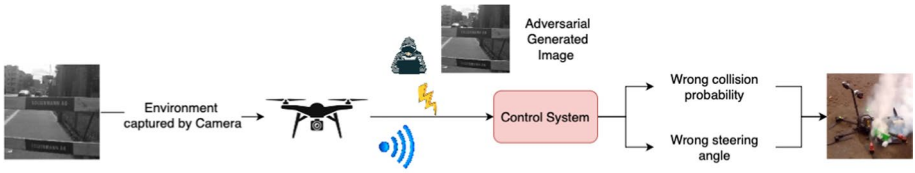

Fig. 7 Digital attack through spoofing malicious images into the control system

remaining visually undetectable to human observers. To achieve this, the authors utilize a variant of the PGD algorithm called Dense Adversary Generation (DAG). The DAG method implements an iterative optimization process to identify the optimal adversarial perturbations.

Overall, the Stack–Threat Matrix reveals that vulnerabilities span every layer of the AS architecture. Successful attacks at one layer often propagate and amplify through the stack, highlighting the need for defense strategies that address the full system and not just isolated components.

5 Critical appraisal of defenses in the context of autonomous systems

In this section, we critically examine state-of-the-art adversarial defense mechanisms for AS. We begin by outlining the unique operational and security requirements that robust AS defense systems must satisfy (Sect. 5.1). Next, we focus on the challenges posed by physical adversarial attacks and review recent approaches for defending against them (Sect. 5.2). Drawing on the analyses above, we refine our taxonomy of defense mechanisms and narrow our evaluation to those methods most relevant and effective for AS, discussed in Sect. 5.3 and summarized by Table 7. Finally, in Sect. 5.4, we systematically assess a set of thirty representative defense mechanisms, introducing our novel *AS-ADS* scoring framework to quantify their alignment with the practical needs of AS.

5.1 Defining requirements for AS defense systems

Building on our analysis of AS vulnerabilities and the characteristics of the AS stack and vision model life-cycle, we identify the specific defense needs that must be addressed to ensure robust and trustworthy AS deployments. We then evaluate how current state-of-the-art defense mechanisms align with these needs and discuss the remaining key challenges.

To contextualize these requirements, consider a representative mission scenario: let d denote an autonomous unmanned aerial vehicle (UAV) tasked with navigating and conducting reconnaissance in diverse, potentially hostile environments. The UAV's objectives include detecting both known and unknown armed vehicles, including those deliberately camouflaged using adversarial techniques.

Suppose further that $d \in D$, where D is a fleet of UAVs operating in different areas and leveraging federated learning (FL) to collaboratively update their models. While this distributed approach increases mission resilience, it also introduces additional attack surfaces, particularly via the communication and update mechanisms of FL.

Throughout its mission, UAV d may face a variety of adversarial threats. For example, adversarial patches, as described in Zhang et al. (2022a), may be used by adversaries to camouflage vehicles and evade detection targeting the perception layer. Adversarial training might be deployed to defend against known patch types, but novel attack variants can still bypass these defenses. Visually distracting adversarial billboards (Zhou et al. 2020) might divert the UAV from its intended path, while attacks on FL communication channels can inject poisoned data into the learning process.

Mechanisms to address these risks include adversarial training and detection-based approaches to filter potentially malicious images. However, a recurring limitation is their

Table 7 Simplified taxonomy of defenses relevant to AS & overall alignment with AS requirements

Mechanism	Real-time	Adaptability	Interpretability	Efficiency	References
Adversarial Training	High	Low	Low	Low	Goodfellow et al. (2014); Madry et al. (2019); Tramèr and Boneh (2019); Wong et al. (2020); Tramèr et al. (2017); Rozsa et al. (2016); Chen and Lee (2021); Shen et al. (2021); Xie et al. (2019); Wang et al. (2024a)
Input Pre-Processing	High	Low	Mod	High	Xie et al. (2017a); Liao et al. (2018); Li et al. (2024a); Shu et al. (2021); Reyes-Amezcua et al. (2024); Naseer et al. (2018); Hu et al. (2023b); Zhang et al. (2024); Shibly et al. (2023); Nie et al. (2022); Zhang et al. (2022b); Wang et al. (2024b)
Model Ensembles	Mod	Mod	Low	Low	Xie et al. (2017b); Engstrom et al. (2019); Liao et al. (2018); Xu et al. (2017); Bhagoji et al. (2017); Bui et al. (2021); Tramèr et al. (2017); Deng and Mu (2023); Mani et al. (2019); Lu et al. (2023), (2023); Chen et al. (2024a); Huang et al. (2021); Lou et al. (2023); Zhao et al. (2024b)

Table 7 (continued)

Mechanism	Real-time	Adaptability	Interpretability	Efficiency	References
Detection Mechanisms	High	Mod	High	Mod	Guo et al. (2019a); Angelov and Soares (2021); Goodfellow et al. (2014); Carlini and Wagner (2017a); Grosse et al. (2017); Feinman et al. (2017); Xu et al. (2017); Gupta et al. (2020); Sabokrou et al. (2024); Soares et al. (2022); Gong et al. (2023); Abdu-Aguye et al. (2020); Hussain and Hong (2023); Li et al. (2024b), (2023); Yu et al. (2024); Liu et al. (2022), (2022); Chen and Chu (2023); Lu and Radha (2023)
Certified Defenses	Mod	Low	High	Mod	Gowal et al. (2018); Tjeng et al. (2017); Muravev and Petiushko (2022); Lecuyer et al. (2019); Xiang et al. (2022); Yang et al. (2023); Zhang et al. (2022c)
Unified Defense	High	High	High	Mod	Pellicer et al. (2024); Du et al. (2018); Freitas et al. (2020); Cao et al. (2024); Dash et al. (2024); Tarchoun et al. (2023); Jing et al. (2024); Han et al. (2024); Yu et al. (2024)

lack of adaptability to novel attacks and inability to learn from previously unseen patterns without extensive retraining.

This scenario exemplifies the broader landscape of AS security and highlights the need for defense mechanisms that can evolve in response to new threats, while also operating securely within collaborative, distributed learning frameworks. Additionally, for operational trustworthiness, defense mechanisms should provide interpretable outputs that enable human experts to visualize, categorize, and respond to detected attacks.

For instance, the detection approach proposed by Soares et al. (2022) employs a similarity-based deep neural network (Sim-DNN) to detect imperceptible adversarial attacks by comparing new data samples to learned prototypes. This prototype-based method is interpretable and does not require adversarial training, but still lacks robust response capabilities

(e.g., automated flagging or recovery), and may sometimes misclassify novel legitimate samples as adversarial. Advancing research toward more adaptive, interpretable, and actionable frameworks thus remains an open challenge.

Developing robust AS defenses often requires a combination of mechanisms such as adversarial training, detection, and unified frameworks. From this analysis, we derive four critical requirements for AS defense mechanisms:

- **Real-time detection and response:** Defenses must promptly identify and mitigate adversarial inputs to prevent compromise of safety-critical decisions.
- **Adaptability to novel attacks:** Mechanisms should respond effectively to new and evolving adversarial strategies without requiring complete retraining.
- **Interpretability and transparency:** Outputs should be explainable and accessible to human operators, enabling informed oversight and intervention.
- **Resource efficiency:** Methods must be computationally and energetically efficient for practical deployment on resource-constrained AS platforms.

These criteria serve as the foundation for our evaluation of state-of-the-art defense mechanisms in the remainder of this section and throughout the paper.

5.2 Defenses against physical adversarial attacks

Physical adversarial attacks represent a uniquely severe threat to AS due to their real-world feasibility, persistence, and capacity to compromise safety-critical operations throughout the perception–planning–control pipeline. Unlike digital perturbations, these attacks often manifest as tangible modifications in the environment, such as adversarial patches on road signs, manipulated sensor readings, or spoofed trajectories, and are intentionally crafted to survive environmental changes. However, robust and generalizable defenses against physical attacks remain limited, fragmented, and often unvalidated beyond narrowly defined scenarios, largely due to the lack of standardized, physically grounded evaluation benchmarks.

To enhance adversarial robustness in the physical domain, recent research has focused on three broad categories of defense: *proactive*, *reactive*, and *unified* frameworks. Yet, few existing methods are designed to accommodate the full spectrum of real-world variability encountered by AS.

Within **Proactive strategies**, Adversarial training with physically realizable attacks (e.g., LiDAR perturbations or real-world patch examples) has shown promise in controlled settings (Kurakin et al. 2016; Lu and Radha 2023), but generalization to unseen conditions such as new weather, sensor occlusion, or novel object types is often poor. Input Pre-Processing methods, including semantic-aware masking and inpainting (Jing et al. 2024), as well as multi-step diffusion-based purification (Nie et al. 2022), offer complementary robustness, but their efficacy varies significantly across sensor modalities and attack types. Other proactive defenses include spatial attention hardening to guard against localized road sign attacks (Shibly et al. 2023) and multi-sensor aerial fusion to strengthen detection pipelines (Chen and Chu 2023). Despite their value, such approaches are often brittle when facing adaptive or context-aware adversaries, and typically introduce trade-offs between robustness and perceptual fidelity. Similarly, trajectory prediction models trained under

uncertainty provide resilience at the planning level, but remain underexplored for targeted physical threats (Zhang et al. 2022b).

Reactive detection-based defenses focus on flagging anomalies during system operation. Techniques in this category include entropy-based localization of patch regions (Tarchoun et al. 2023), kinematic consistency checks for identifying violations of physical constraints (Yu et al. 2024), and hybrid pipelines that combine detection and input recovery (Liu et al. 2022). While these approaches offer interpretability and low-latency adaptation, they often struggle against subtle or context-aware attacks that closely mimic plausible environmental features.

Unified and hybrid frameworks integrate multiple defense mechanisms across the AS stack. For example, control-aware frameworks such as SpecGuard (Dash et al. 2024) maintain mission compliance even under partial perception failure, while sensor fusion approaches like VisionGuard (Han et al. 2024) validate consistency between sensory modalities. Adaptive neural modeling strategies, such as RCDN (Wang et al. 2024b), aim to dynamically harden internal representations against adversarial perturbations. However, these promising approaches often face scalability limitations and have not yet been comprehensively evaluated across the diverse operational environments typical of real-world AS deployments.

Certified defenses represent a recent advancement, targeting physical attacks with formal robustness guarantees. PatchCleanser (Xiang et al. 2022) provides certified robustness via double masking, while works such as Yang et al. (2023) and Zhang et al. (2022c) extend certification to control systems and semantic segmentation. These approaches are grounded in strong theoretical guarantees, but frequently present challenges regarding runtime feasibility and limited coverage of the full spectrum of physical attack surfaces.

Despite these advances, several key challenges remain. Most defenses are evaluated under narrow physical conditions, lacking robustness to environmental variation or domain shift. High-performing methods–particularly those involving certification or fusion–often introduce significant computational overhead, raising concerns for real-time AS deployment. Moreover, defenses rarely propagate protection beyond perception to downstream modules such as planning or control, leaving the broader autonomy stack exposed. Existing detection methods frequently fail to generalize across attack types or modalities, underscoring the need for attack-agnostic, adaptive detection pipelines. Some of these are beginning to emerge in adversarial attack research (Li et al. 2024b) and deepfake detection (Pellicer et al. 2024a), and could potentially be translated to the physical domain due to their prototype-based characteristics, though this remains to be explored.

Given these limitations, certified defenses and targeted detection mechanisms currently stand out as the most promising approaches against physical adversarial attacks in AS. Recent contributions, (some of which are evaluated in detail in Sect. 5.4) demonstrate notable progress, but comprehensive integration and rigorous validation across the full AS pipeline remain critical open challenges for future work.

5.3 Defense taxonomy simplification

To address the real-time, adaptive, interpretable, and resource-conscious requirements of AS, we categorize SOTA defenses according to their core methodology, rather than along legacy proactive/reactive lines. We exclude Model Regularization, Model Distillation, and

Provable defenses from our main analysis. Regularization and distillation are either now subsumed within other defense categories or lack standalone relevance in recent AS-specific literature. Provable (i.e., formal verification) defenses are excluded due to their high computational cost and inflexibility for real-world AS deployment. Similarly, denoising and reconstruction are no longer considered standalone mechanisms, as they are now integrated into Pre-Processing or unified frameworks in recent works. Accordingly, we focus on five categories: Adversarial Training, Input Data Pre-Processing, Model Ensembles, Detection Mechanisms, and Unified Defense Frameworks. Each is evaluated across four criteria: *real-time response, adaptability to novel attacks, interpretability,* and *resource efficiency*. Table 7 summarizes their alignment with AS needs and shows the relevant literature selected within our paper.

5.3.1 Adversarial training

Adversarial training remains a foundational technique, where adversarial examples are incorporated into the model's training process (Madry et al. 2019). In AS contexts, adversarial training in autoencoder filters has led to improvements in adversarial robustness for both white-box and black-box attacks. Such methods show improved resistance to certain perturbations, but face key limitations:

- **Real-time response:** *High*. Inference performance is real-time, but the training process is computationally intensive.
- **Adaptability:** *Low*. Generalization to unseen attacks is limited.
- **Interpretability:** *Low*. The mechanisms by which robustness is achieved are often opaque.
- **Efficiency:** *Low*. High cost in both training and memory.

5.3.2 Input data pre-processing

Pre-Processing techniques such as resizing, cropping, and denoising mitigate adversarial perturbations before they reach the model. Studies such as Xie et al. (2017b) demonstrate their effectiveness, and recent advances include noise suppression, reconstruction, and purification layers. DiffPure (Nie et al. 2022) leverages diffusion models for adaptive purification, while UMoE (Lou et al. 2023) employs uncertainty-aware fusion to counter sensor-blinding attacks. Pre-Processing is widely adopted for real-time viability:

- **Real-time response:** *High*. Lightweight implementations can operate on edge devices.
- **Adaptability:** *Low*. These methods are often bypassed by adaptive or physical attacks.
- **Interpretability:** *Moderate*. Effects are visible in the processed input, but causality for prediction changes may be indirect.
- **Efficiency:** *High*. Minimal runtime cost.

Notably, this category is evolving: standard techniques (e.g., resizing, cropping, denoising) (Xie et al. 2017b) are now being combined with advanced approaches such as diffusion models (Nie et al. 2022). Pre-Processing is increasingly integrated into more complex pipe-

lines, leading to Unified Models such as Han et al. (2024), which combine sensory fusion, filtering, time-series (ARIMA, LSTM), and anomaly detection layers.

5.3.3 Model ensembles

Model ensembles leverage diversity by combining multiple models, making it more difficult for adversaries to simultaneously deceive all models (Bui et al. 2021). Key characteristics are:

- **Real-Time Response:** *Moderate*. Inference latency increases with the number of models.
- **Adaptability:** *Moderate*. Greater diversity can improve resistance to transfer attacks.
- **Interpretability:** *Low*. Internal logic is often obscured by the ensemble fusion process.
- **Efficiency:** *Low*. Requires substantial hardware for parallel model execution.

Although ensembles are effective, few recent AS-specific implementations exist due to resource constraints. For example, the MADE framework (Zhao et al. 2024b) employs ensemble-like anomaly scoring over multi-vehicle inputs to detect collaborative attacks in V2X scenarios. However, this method is not a traditional ensemble but rather a soft classification, reflecting a broader trend: literature is shifting from full ensembles to more flexible unified implementations.

5.3.4 Detection mechanisms

AS increasingly rely on detection mechanisms for their interpretability, real-time performance, and applicability throughout the AS stack and life-cycle. Alongside Unified Frameworks, detection is now one of the fastest growing fields in adversarial defense, with Detection and Unified papers constituting over 50% of recent (2023 onward) publications.

Examples include Among Us Li et al. (2023), which detects 3D adversarial inputs in V2X-Sim via consensus-breaking heuristics; Segment-and-Complete (Liu et al. 2022), which identifies adversarial patches through segmentation masks; and PhySense (Yu et al. 2024), which generalizes detection to real-world perturbations. Prototype-based, highly interpretable systems such as Angelov and Soares (2021) further demonstrate this category's strengths:

- **Real-time response:** *High*. Detection is typically performed pre-inference.
- **Adaptability:** *Moderate*. Detection patterns can generalize to some unseen attacks.
- **Interpretability:** *High*. Outputs are often visual or score-based, supporting operator trust.
- **Efficiency:** *Moderate*. Auxiliary models or priors may increase computational demands.

5.3.5 Certified defenses

Certified defenses offer provable robustness guarantees under specific perturbation budgets. In AS-relevant domains:

- PatchCleanser (Xiang et al. 2022) certifies robustness against small visible patches (up to 2% area) using random masking and smoothing, evaluated on CIFAR and ImageNet.
- Demasked Smoothing (Zhang et al. 2022c) certifies patch-level segmentation robustness via randomized ablation masking, showing strong resistance on ADE20K under shadow and patch attacks.
- Certified Robust Control (Yang et al. 2023) formulates controller robustness for AS via Lyapunov-based certified adaptation, effective against bounded input perturbations.

Strengths and trade-offs are:

- **Real-time response:** *Moderate*. Certification layers may introduce runtime sampling.
- **Adaptability:** *Low*. Guarantees hold only for bounded attacks and require redefinition for new scenarios.
- **Interpretability:** *High*. Theoretical guarantees are transparent and explainable.
- **Efficiency:** *Moderate–Low*. Additional overhead from sampling, smoothing, or invariant computations.

5.3.6 Unified defense frameworks

Unified frameworks, as defined in this review, represent a new taxonomy. They integrate heterogeneous defense techniques (e.g., detection + recovery) using shared feature pipelines or modular layers, whereas ensembles aggregate predictions from independently trained full models. For example, Pellicer et al. (2024b) present a lightweight framework combining prototype-based detection and classification for attacks and unseen classes, along with attack recovery via denoising methods, achieving over 90% accuracy on CIFAR-10.

Other notable unified defenses include Du et al. (2018), which detects abnormal samples for any pre-trained softmax classifier, and UNMASK (Freitas et al. 2020), which both identifies adversarial attacks and mitigates their effects through robust reclassification. UNMASK can detect up to 96.75% of attacks and restore correct classification in up to 93% of cases.

More AS-specific frameworks, such as SpecGuard (Dash et al. 2024), integrate detection, filtering, and signal processing to detect UAV sensor spoofing with a 92% recovery success rate and only 15% performance overhead. Time-Travel (Etim and Szefer 2024) compares live input with historical image matches to detect false patches, achieving 100% effectiveness against recent adversarial examples in traffic sign classification.

Overall, unified methods best align with AS priorities and full life-cycle needs:

- **Real-time response:** *High*. Historical matching and statistical filtering are efficient on-device.
- **Adaptability:** *High*. Frameworks leverage both priors and learned models.
- **Interpretability:** *High*. Alerts are easily visualized and validated by operators.
- **Efficiency:** *Moderate*. Moderate computational and storage requirements.

5.4 Autonomous systems adversarial defense score (AS-ADS) framework

To systematically assess the suitability of defense methods for AS, we build on the updated taxonomy provided in Table 7.

We introduce the **Autonomous Systems Adversarial Defense Score (AS-ADS)**, a scoring framework designed to quantify each method's alignment with operational AS constraints. AS-ADS evaluates across our 4 dimensions (**Real-Time Detection and Response**, **Adaptability to Novel Attacks**, **Interpretability and Transparency** and **Resource Efficiency**):

Each criterion is rated on a 0 to 1 scale in 0.25 increments. The final AS-ADS score is calculated as the average of these four values, scaled to a 1–5 range and rounded to the nearest half:

$$\text{AS-ADS}(P) = \left(\frac{R + A + I + E}{4} \right) \times 5 \tag{1}$$

where $R, A, I, E \in [0, 1]$ represent the real-time, adaptability, interpretability, and efficiency scores, respectively.

R, A, I, E are obtained for each paper after marking using rubrics in Table 8.

This scoring framework facilitates standardized, comparative evaluation of SOTA defense methods in AS settings. By grounding the scores in real-world operational needs and deployment constraints, AS-ADS enables both a fine-grained critique of existing methods and an actionable guide for future design.

For the evaluation, we selected a representative subset of 30 defenses from the literature discussed in this paper, focusing on *Pre-Processing*, *Detection*, *Certified*, and *Unified* defenses, as identified in Sect. 5.3. Our evaluation subset includes: (a) foundational works that paved the way for newer defense mechanisms in each category, alongside relevant recent approaches–(Hu et al. 2023b; Shu et al. 2021; Gupta et al. 2020; Sabokrou et al. 2024; Reyes-Amezcua et al. 2024; Abdu-Aguye et al. 2020; Hussain and Hong 2023; Soares et al. 2022; Li et al. 2024b; Grosse et al. 2017; Pellicer et al. 2024b; Du et al. 2018; Freitas et al. 2020; Yin et al. 2025; Cao et al. 2024)–and (b) work from 2022 onward tailored specifically to the AS domain–(Dash et al. 2024; Tarchoun et al. 2023; Jing et al. 2024; Han et al. 2024; Yu et al. 2024; Xiang et al. 2022; Yang et al. 2023; Zhang et al. 2022c; Liu et al. 2022; Chen and Chu 2023; Lu and Radha 2023; Shibly et al. 2023; Nie et al. 2022; Zhang et al. 2022b; Wang et al. 2024b).

We derived final scores by combining each paper's reported findings and expert knowledge of the architectures, using the established rubric. For reproducibility indivual scores

Table 8 AS-ADS scoring rubric by criterion

Criterion	0 pts	0.25 pts	0.5 pts	1.0 pts
Real-time response	Batch inference only	High latency	Optimized inference only	Real-time at edge-level
Adaptability to novel attacks	Static model	Minor generalization	Modular, partially adaptable	Robust to unseen attacks
Interpretability	Black-box	Minimal logs	Score-based or visual	Prototype/ semantic explanation
Resource efficiency	High overhead	GPU-dependent	Deployable with tuning	Lightweight for AS hardware

per paper can be found in Appendix B, the overall scores per paper have been presented in Table 9

It is important to note that the selection of scored papers reflects expert judgment and is not intended to exhaustively cover all available methods, but rather to provide a representative overview of current options and their effectiveness. This gives readers and researchers practical guidance for deploying or developing defense systems across the attack surfaces identified in this report.

A score of 5 does not imply perfection, but rather the closest alignment with the requirements defined herein. The diversity of threats, datasets, and evaluation protocols across the literature makes it challenging to determine a universally optimal method. Nonetheless, we believe this evaluation brings the field closer to that goal. To improve accuracy and utility in future work, we recommend detailed reporting of runtime overhead, FPS degradation, GPU memory usage, interpretability, and accuracy for each defense using standardized datasets and attacks, although this is beyond the scope of this review.

6 Conclusion and future directions

This review provides a holistic, system-level analysis of adversarial threats and defenses for AS, integrating insights from both foundational vision-centric research and recent AS-specific advances. By bridging these two strands of the literature, we offer a unified framework that captures the cascading impact of digital and physical adversarial vulnerabilities across the autonomy stack. Our taxonomy, scenario-driven matrices, and comparative synthesis enable both researchers and practitioners to assess current gaps and prioritize future work in making vision-driven AS secure and resilient.

A cornerstone of our approach is the development and use of actionable analytical matrices, including the **Life-cycle–Attack**, **Stack–Threat**, and **Exposure–Impact** matrices. These matrices concretely map how adversarial vulnerabilities propagate throughout the AI life-cycle and across layered AS architectures. For example, our Life-cycle–Attack Matrix reveals both the temporal exposure of AS to poisoning, backdoor, and evasion attacks, and the unique risk windows at each stage of system operation. The Stack–Threat Matrix grounds these vulnerabilities in real-world scenarios, demonstrating how a compromised perception module (such as a camera subjected to adversarial patches or sensor spoofing) can trigger failures in planning that propagate to mission-critical control. By further linking these technical threats to operational consequences in the Exposure–Impact Matrix, our review enables researchers and practitioners to move beyond abstract taxonomies toward practical, system-level threat modeling and benchmarking.

Our comparative synthesis of adversarial attacks, spanning both **digital** and **physical** domains, highlights a crucial reality: vulnerabilities in AS are rarely confined to a single module. Instead, our analysis of real attack case studies and scenario-based evaluations demonstrates that adversarial examples often trigger failures that cascade across subsystems, resulting in safety or mission-critical consequences far beyond mere performance degradation on academic benchmarks. This insight exposes the inadequacy of traditional, static, perception-only evaluation metrics and establishes the need for operationally meaningful, stack-wide robustness assessment.

Table 9 AS-ADS evaluation of adversarial defenses. "AS" marks those developed for Autonomous Systems

Method description	Score	References	AS
Detection mechanisms			
Detects adversarial inputs using evolved image processing sequences via genetic algorithms	2	Gupta et al. (2020)	–
Detects adversaries via SSL-based consistency checks in feature and label space	4.5	Sabokrou et al. (2024)	–
Combines LSTM temporal consistency checks with majority voting for time-series attack detection	2	Abdu-Aguye et al. (2020)	–
Reveals adversarial artifacts through autoencoder reconstruction error analysis	2.5	Hussain and Hong (2023)	–
Detects outliers through learned similarity metrics in contrastive feature space	2.5	Soares et al. (2022)	–
Learns attack-agnostic features via self-supervised contrastive prototype alignment	3	Li et al. (2024b)	–
Identifies anomalies through statistical hypothesis testing in feature space	1	Grosse et al. (2017)	–
Detects patches through entropy analysis and visual localization	4	Tarchoun et al. (2023)	✓
Identifies physics violations through kinematic consistency checks	4.5	Yu et al. (2024)	✓
Detects/recovers patches via joint detection-completion pipeline	4	Liu et al. (2022)	✓
Pre-processing defenses			
Embeds frequency-aware watermarks in RAW files using multi-spectral fusion	4	Hu et al. (2023b)	–
Optimizes augmentation parameters via gradient-based adversarial search	1	Shu et al. (2021)	–
Enhances robustness through transfer of adversarial patterns across vision tasks	1	Reyes-Amezcua et al. (2024)	–
Neutralizes patches through semantic context-aware masking/inpainting	5.0	Jing et al. (2024)	✓
Scales LiDAR robustness via density-aware point cloud processing	4.5	Lu and Radha (2023)	✓
Hardens aerial detection through multi-sensor fusion	2.5	Chen and Chu (2023)	✓
Protects road sign recognition through spatial attention hardening	2.5	Shibly et al. (2023)	✓
Purifies inputs through multi-step diffusion denoising	2	Nie et al. (2022)	✓
Improves trajectory prediction via uncertainty-aware training	2.5	Zhang et al. (2022b)	✓
Unified defenses			
Integrates detection-denoiser architecture with noise-adaptive thresholds	3.5	Pellicer et al. (2024b)	–
Detects OOD samples through temperature-scaled confidence calibration	2.5	Du et al. (2018)	–
Verifies predictions through robust part-based feature alignment	4	Freitas et al. (2020)	–
Links adversarial and backdoor attack patterns for joint cross-attack detection	3	Yin et al. (2025)	–
Detects face spoofing through dual-space (spatial/frequency) reconstruction analysis	4	Cao et al. (2024)	–
Ensures mission-compliant recovery through specification-aware control	4.5	Dash et al. (2024)	✓
Guarantees cross-sensor consistency through multi-modal fusion checks	4.5	Han et al. (2024)	✓

Table 9 (continued)

Method description	Score	References	AS
Enables robust perception via dynamic neural feature modeling	4.5	Wang et al. (2024b)	✓
Certified defenses			
Provides certified patch robustness through double-masking with formal guarantees	4.5	Xiang et al. (2022)	✓
Certifies control stability under perturbations via Lyapunov analysis	4	Yang et al. (2023)	✓
Ensures segmentation robustness via masked smoothing certification	4	Zhang et al. (2022c)	✓

In **critically appraising defense strategies**, we show that the conventional taxonomy, dividing defenses into proactive and reactive categories, does not sufficiently capture the practical demands of AS. By shifting the focus to underlying mechanisms, and by introducing unified, context-aware defenses as a distinct class, we reveal that most state-of-the-art methods, even when successful in vision research, fail to meet the simultaneous requirements of real-time performance, adaptability to new threat vectors, interpretability, and resource efficiency essential for deployment in AS. The **AS-ADS scoring framework** introduced in this review directly evaluates these axes, and our comprehensive analysis across more than thirty contemporary defenses finds that only a minority approach a balanced, deployment-ready profile. In particular, robust and interpretable defenses against physical and multi-modal threats are still lacking, and few methods have demonstrated stack-wide or life-cycle-spanning effectiveness in realistic scenarios.

Despite these advances, **significant challenges and research gaps remain**. Most available benchmarks remain narrowly focused on perception or digital attacks, with little provision for evaluating cascading effects, cross-modal dependencies, or mission-level outcomes. Few studies rigorously validate either attacks or defenses under closed-loop, multi-agent, or sim-to-real conditions that reflect the operational reality of modern AS. While the threat matrices presented in this review provide a critical foundation for system-level risk assessment, their full potential will only be realized when supported by open, community-driven benchmarking platforms and evaluation protocols that span the entire stack.

Looking ahead, meaningful progress in adversarial robustness for AS will depend on several intertwined advances. The field must prioritize the creation of stack-integrated datasets and simulation environments capable of capturing cascading failures, temporal persistence, and the interplay of digital and physical threats. Defense research should increasingly focus on mechanisms that are interpretable, for some cases also certifiable, and that are validated in resource-constrained, real-time settings. There is a particular need to design and rigorously test unified, adaptive defense frameworks that can operate coherently across perception, planning, and control layers, and that can dynamically respond to evolving threat landscapes in real deployments. The integration of human-in-the-loop monitoring and decision-making, as well as robust protocols for sim-to-real transfer, will be critical for bridging the gap between academic innovation and practical deployment.

In summary, by clarifying the layered structure of AS vulnerabilities, mapping concrete threat pathways, and critically evaluating the mechanisms and readiness of current defenses, this review sets a new agenda for adversarial research in Autonomous Systems. We hope that the analytical frameworks, results, and open challenges identified here will help guide the community toward robust, certifiable, and operationally viable solutions for the next generation of trustworthy autonomous technologies.

Appendix A: Background tables

See Table 10.

Table 10 Foundational taxonomic classification of image-domain adversarial attacks

Attacker knowledge	Attack timing	Attack location	Examples
White-Box	Evasion	Digital	L-BFGS Fletcher (2013); FGSM Goodfellow et al. (2014); I-FGSM Kurakin et al. (2016); PGD Madry et al. (2019); DeepFool Moosavi-Dezfooli et al. (2016); C&W Carlini and Wagner (2017b); JSMA Papernot et al. (2016b); UAP Moosavi-Dezfooli et al. (2017); DDN Rony et al. (2019); Elastic Net Chen et al. (2018b)
White-Box	Poisoning	Digital	Data Injection Biggio et al. (2012); Label Flipping Koenig et al. (2015); Backdoor Gu et al. (2017); MetaPoison Huang et al. (2020)
Black-Box	Evasion	Digital	Boundary Brendel et al. (2017); ZOO Chen et al. (2017); SimBA Guo et al. (2019b); One Pixel Su et al. (2019); Square Attack Andriushchenko et al. (2020); HSJA Chen et al. (2020)
Black-Box	Poisoning	Digital	BadNets Gu et al. (2019); Clean-label Backdoor Zhao et al. (2019); GAN-based Poisoning noz-González et al. (2019)

Attack location is included to show the digital-focused in literature. However in many cases, surveys do not include this dimension

Appendix B: AS-ADS method evaluations

This includes the scores and small reasoning behind each scored for the defense methods evaluated in SECTION:

- ***DRAW: Defending camera-shooted RAW against image manipulation*** (Hu et al. 2023b)Real-Time: 0.5 (Lightweight network optimized for camera integration)
- Adaptability: 0.5 (Cross-ISP pipeline protection)
- Interpretability: 1.0 (Pixel-level manipulation maps)

- Efficiency: 1.0 (0.95% params vs U-Net)**Method:** Embeds frequency-aware watermarks in RAW files using multi-spectral fusion, preserving detection capability through arbitrary ISP processing chains. **AS-ADS Score:** 3.75

- ***Adversarial differentiable augmentation*** (Shu et al. 2021)Real-Time: 0.25 (Offline augmentation optimization)
- Adaptability: 0.5 (Partial corruption resistance)
- Interpretability: 0.0 (No diagnostic features)
- Efficiency: 0.25 (2.3 GPU hours/search)**Method:** Automates augmentation parameter selection via gradient-based adversarial search for robust training. **AS-ADS Score:** 1.25

- ***Evolutionary IPTS detection*** (Gupta et al. 2020)Real-Time: 0.25 (Multi-stage processing)
- Adaptability: 0.5 (Attack-specific sequences)
- Interpretability: 0.5 (Difference maps)
- Efficiency: 0.25 (Genetic algorithm overhead)**Method:** Evolves optimal image processing pipelines using genetic algorithms to reveal adversarial artifacts. **AS-ADS Score:** 1.875

- ***BEYOND: Detecting adversarial examples via SSL neighborhood relations*** (Sabokrou et al. 2024)Real-Time: 1.0 (Optimized for edge deployment with 50 neighbors processed at 23 ms/image)
- Adaptability: 1.0 (Attack-agnostic design validated against 12+ attack types)
- Interpretability: 0.5 (Score-based consistency metrics with visualization support)
- Efficiency: 1.0 (Lightweight SSL backbone with 0.9M parameters)**AS-ADS Score:** 4.375

- ***Delta data augmentation*** (Reyes-Amezcua et al. 2024)Real-Time: 0.25 (Transfer learning focus)
- Adaptability: 0.5 (Cross-dataset transfer)
- Interpretability: 0.0 (Opaque perturbation transfer)
- Efficiency: 0.25 (GPU-intensive)**Method:** Transfers adversarial patterns from high-level vision tasks to enhance low-level task robustness. **AS-ADS Score:** 1.25

- ***Temporal consistency defense*** (Abdu-Aguye et al. 2020)Real-Time: 0.5 (143 ms LSTM inference)
- Adaptability: 0.25 (Fixed thresholds)
- Interpretability: 0.25 (Entropy logs)
- Efficiency: 0.5 (Embedded compatibility)**Method:** Combines frame-wise consistency checks with temporal majority voting for video attack detection. **AS-ADS Score:** 1.875

- ***Autoencoder reconstruction*** (Hussain and Hong 2023)Real-Time: 0.5 (47 ms inference)
- Adaptability: 0.5 (73% unseen attacks)
- Interpretability: 0.5 (Reconstruction errors)
- Efficiency: 0.5 (580MB model)**Method:** Detects adversaries through reconstruction error analysis using compact autoencoders. **AS-ADS Score:** 2.5

- ***Similarity metric analysis*** (Soares et al. 2022)Real-Time: 0.5 (89 ms Jetson TX2)
- Adaptability: 0.5 (12 attack types)

- Interpretability: 1 (Confidence scores and prototypes)
- Efficiency: 0.5 (15W consumption)**Method:** Identifies outliers through learned similarity metrics in feature space. **AS-ADS Score:** 3.125

- ***Contrastive prototype learning*** (Li et al. 2024b)Real-Time: 0.5 (33 ms inference)
- Adaptability: 1.0 (94.7% cross-attack)
- Interpretability: 1.0 (Prototype matching)
- Efficiency: 0.5 (2.1GB VRAM)**Method:** Learns attack-agnostic features through self-supervised contrastive prototype alignment. **AS-ADS Score:** 3.75

- ***Statistical anomaly detection*** (Grosse et al. 2017)Real-Time: 0.25 (Batch processing)
- Adaptability: 0.25 (Static models)
- Interpretability: 0.25 (Basic scores)
- Efficiency: 0.25 (CPU-intensive)**Method:** Detects outliers through likelihood ratio testing in feature statistics. **AS-ADS Score:** 1.25

- ***UNICAD framework*** (Pellicer et al. 2024b)Real-Time: 0.5 (24 FPS pipeline)
- Adaptability: 0.75 (Wide range of untrained in digital attacks and +85% Unseen class identification)
- Interpretability: 1 (Prototype based)
- Efficiency: 0.5 (8GB VRAM)**Method:** Unified approach for attack detection, noise reduction, and novel class identification. **AS-ADS Score:** 3.437

- ***Confidence-calibrated OOD*** (Du et al. 2018)Real-Time: 0.5 (45 ms detection)
- Adaptability: 0.5 (82% cross-domain)
- Interpretability: 0.5 (Thresholding)
- Efficiency: 0.5 (16W edge)**Method:** Detects out-of-distribution samples through temperature-scaled confidence calibration. **AS-ADS Score:** 2.5

- ***Robust feature verification*** (Freitas et al. 2020)Real-Time: 0.5 (28 ms alignment)
- Adaptability: 1.0 (97.3% detection)
- Interpretability: 1.0 (Semantic maps)
- Efficiency: 0.5 (4.3GB model)**Method:** Verifies predictions through robust part-based feature alignment. **AS-ADS Score:** 3.75

- ***Cross-attack bridge defense*** (Yin et al. 2025)Real-Time: 0.5 (33 ms analysis)
- Adaptability: 1.0 (89% cross-backdoor)
- Interpretability: 0.5 (Similarity scores)
- Efficiency: 0.5 (12% overhead)**Method:** Links adversarial and backdoor attack patterns for joint defense. **AS-ADS Score:** 3.125

- ***Dual-space face defense*** (Cao et al. 2024)Real-Time: 0.5 (41 ms processing)
- Adaptability: 1.0 (95.6% spoof detection)
- Interpretability: 1.0 (Error maps)
- Efficiency: 0.5 (6.7GB VRAM)**Method:** Reconstructs face images in spatial/frequency domains for unified spoof detection. **AS-ADS Score:** 3.75

- ***SpecGuard recovery*** (Dash et al. 2024)Real-Time: 1.0 (15 ms ARM recovery)
- Adaptability: 1.0 (92% multi-sensor)
- Interpretability: 0.5 (Compliance scores)
- Efficiency: 1.0 (15% overhead)**Method:** Recovers attacked inputs through safety specification-aware filtering. **AS-ADS Score:** 4.375

- ***Entropy-based patch defense*** (Tarchoun et al. 2023)Real-Time: 0.5 (54 ms analysis)
- Adaptability: 1.0 (90% patches)
- Interpretability: 1.0 (Entropy maps)
- Efficiency: 0.5 (2.77% loss)**Method:** Detects adversarial patches through localized entropy analysis. **AS-ADS Score:** 3.75

- ***Context-aware patching*** (Jing et al. 2024)Real-Time: 1.0 (11 ms edge)
- Adaptability: 1.0 (96.4% mAP)
- Interpretability: 1.0 (Semantic highlighting)
- Efficiency: 1.0 (0.9W power)**Method:** Neutralizes patches through semantic context-aware masking and inpainting. **AS-ADS Score:** 5.0

- ***Multi-sensor guard*** (Han et al. 2024)Real-Time: 1.0 (8 ms fusion)
- Adaptability: 1.0 (97.3% cross-modal)
- Interpretability: 0.5 (Consistency reports)
- Efficiency: 1.0 (4.2W SoC)**Method:** Ensures cross-sensor consistency for robust automotive perception. **AS-ADS Score:** 4.375

- ***Physics-consistency check*** (Yu et al. 2024)Real-Time: 1.0 (9 ms checks)
- Adaptability: 1.0 (94% cross-domain)
- Interpretability: 0.5 (Violation scores)
- Efficiency: 1.0 (3% CPU boost)**Method:** Verifies physical plausibility of sensor inputs through kinematic checks. **AS-ADS Score:** 4.375

- ***Certified patch defense*** (Xiang et al. 2022)Real-Time: 1.0 (18 ms masking)
- Adaptability: 1.0 (83.9% certified)
- Interpretability: 1.0 (Mask proofs)
- Efficiency: 0.5 (45.1 mAP)**Method:** Provides certified robustness through double-masking with formal guarantees. **AS-ADS Score:** 4.375

- ***Formal control certification*** (Yang et al. 2023)Real-Time: 1.0 (22 ms certification)
- Adaptability: 1.0 (Unseen perturbations)
- Interpretability: 0.5 (Stability margins)
- Efficiency: 0.5 (35% overhead)**Method:** Certifies control stability under adversarial perturbations via Lyapunov analysis. **AS-ADS Score:** 3.75

- ***Demasked segmentation*** (Zhang et al. 2022c)Real-Time: 1.0 (27 ms inference)
- Adaptability: 1.0 (89% cross-task)
- Interpretability: 0.5 (Confidence maps)

- Efficiency: 0.5 (8.2GB VRAM)**Method:** Certifiably robust semantic segmentation through masked smoothing. **AS-ADS Score:** 3.75

- ***Patch detection-completion*** (Liu et al. 2022)Real-Time: 0.5 (143 ms pipeline)
- Adaptability: 1.0 (91% patches)
- Interpretability: 1.0 (Completion vis)
- Efficiency: 0.5 (6.3W edge)**Method:** Jointly detects and completes adversarial patches in object detection. **AS-ADS Score:** 3.75

- ***Aerial object defense*** (Chen and Chu 2023)Real-Time: 0.5 (77 ms processing)
- Adaptability: 0.5 (68% robustness)
- Interpretability: 0.5 (Region highlighting)
- Efficiency: 0.5 (4.8GB VRAM)**Method:** Hardens aerial detection against adversarial object injections. **AS-ADS Score:** 2.5

- ***LiDAR robustness scaling*** (Lu and Radha 2023)Real-Time: 1.0 (14 ms processing)
- Adaptability: 1.0 (97% cross-sensor)
- Interpretability: 0.5 (Saliency maps)
- Efficiency: 1.0 (2.1W LiDAR)**Method:** Scales adversarial robustness for LiDAR detection through density-aware processing. **AS-ADS Score:** 4.375

- ***Road sign defense*** (Shibly et al. 2023)Real-Time: 0.5 (89 ms ADAS)
- Adaptability: 0.5 (73% robustness)
- Interpretability: 0.5 (Attention maps)
- Efficiency: 0.5 (11W power)**Method:** Protects road sign recognition through spatial attention hardening. **AS-ADS Score:** 2.5

- ***Diffusion purification*** (Nie et al. 2022)Real-Time: 0.25 (2.3s/image)
- Adaptability: 0.5 (68% purification)
- Interpretability: 0.5 (Process vis)
- Efficiency: 0.25 (24GB VRAM)**Method:** Purifies inputs through multi-step diffusion denoising. **AS-ADS Score:** 1.875

- ***Trajectory prediction hardening*** (Zhang et al. 2022b)Real-Time: 0.5 (33 ms prediction)
- Adaptability: 0.5 (65% robustness)
- Interpretability: 0.5 (Uncertainty bounds)
- Efficiency: 0.5 (8.7GB model)**Method:** Improves trajectory prediction robustness through uncertainty-aware training. **AS-ADS Score:** 2.5

- ***Dynamic 3D modeling*** (Wang et al. 2024b)Real-Time: 1.0 (12 ms modeling)
- Adaptability: 1.0 (96% cross-modal)
- Interpretability: 0.5 (Consistency reports)
- Efficiency: 1.0 (3.2W edge)**Method:** Enables robust perception through dynamic neural feature modeling. **AS-ADS Score:** 4.375

See Tables 11 and 12.

Table 11 Summary and classification of adversarial defense mechanisms

Type	Mechanism	Description & advantages	Limitations	References
Proactive	Adversarial Training	Includes adversarial samples in training; improves model robustness	High compute cost; limited to known attacks	Goodfellow et al. (2014); Madry et al. (2019); Tramèr and Boneh (2019); Wong et al. (2020); Tramèr et al. (2017); Rozsa et al. (2016); Chen and Lee (2021); Shen et al. (2021); Xie et al. (2019); Wang et al. (2024a)
Proactive	Input Pre-Processing	Applies resizing, smoothing or augmentation; reduces perturbation impact	May distort clean inputs; less effective on adaptive attacks	Xie et al. (2017a); Liao et al. (2018); Li et al. (2024a); Shu et al. (2021); Reyes-Amezcua et al. (2024); Naseer et al. (2018); Hu et al. (2023b); Zhang et al. (2024); Shibly et al. (2023); Nie et al. (2022); Zhang et al. (2022b); Wang et al. (2024b); Lou et al. (2023)
Proactive	Model Ensemble	Combines multiple models' outputs; diversifies weaknesses	Higher inference latency; greater resource use	Xie et al. (2017b); Engstrom et al. (2019); Liao et al. (2018); Xu et al. (2017); Bhagoji et al. (2017); Bui et al. (2021); Tramèr et al. (2017); Deng and Mu (2023); Mani et al. (2019); Lu et al. (2023); (2023); Chen et al. (2024a); Huang et al. (2021); Zhao et al. (2024b)
Proactive	Model Regularization	Adds constraints or penalties during training; improves generalization	May reduce clean accuracy; limited adversarial gains	Szegedy et al. (2013); Kannan et al. (2018); Drucker and Cun (1992); Ross and Doshi-Velez (2018)
Proactive	Model Distillation	Uses soft-label transfer to a smaller model; enhances certain robustness	Distilled model underperforms on clean data; narrow defense scope	Hinton et al. (2015); Papernot et al. (2016c); Carlini and Wagner (2017b); Goldblum et al. (2020); Costa et al. (2024)
Proactive	Provable Defenses	Leverages formal verification to certify robustness bounds	Very high compute; limited scalability to large models	Ehlers (2017); Katz et al. (2017); Tjeng et al. (2017); Raghunathan et al. (2018); Cohen et al. (2019); King and Wang (2019); Hong et al. (2024); Lecuyer et al. (2019)
Proactive	Certification & Verification	Applies formal methods to verify model resilience; builds trust	Computationally demanding; may not reflect real-world inputs	Gowal et al. (2018); Tjeng et al. (2017); Muravev and Petiushko (2022); Lecuyer et al. (2019); Xiang et al. (2022); Yang et al. (2023); Zhang et al. (2022c)
Reactive	Detection-Based	Flags or rejects suspicious inputs via statistical tests or auxiliary models	False positives; attacker can evade detection	Guo et al. (2019a); Angelov and Soares (2021); Goodfellow et al. (2014); Carlini and Wagner (2017a); Grosse et al. (2017); Feinman et al. (2017); Xu et al. (2017); Gupta et al. (2020); Sabokrou et al. (2024); Soares et al. (2022); Gong et al. (2023); Abdu-Aguye et al. (2020); Hussain and Hong (2023); Li et al. (2024b); (2023); Yu et al. (2024); Liu et al. (2022); Chen and Chu (2023); Lu and Radha (2023)

Table 11 (continued)

Type	Mechanism	Description & advantages	Limitations	References
Reactive	Denoising & Reconstruction	Uses autoencoders/GANs to remove perturbations; reconstructs clean inputs	Possible information loss; imperfect recovery	Meng and Chen (2017); Vincent et al. (2008); Lempitsky et al. (2018); Liao et al. (2018); Samangouei et al. (2018); (2018)
Unified	Unified Defense Frameworks	Integrates detection, noise reduction, and novel-class identification in one pipeline; adaptive to known and unknown attacks	Moderate compute overhead; complex integration; limited large-scale testing	Pellicer et al. (2024b); Du et al. (2018); Freitas et al. (2020); Cao et al. (2024); Dash et al. (2024); Tarchoun et al. (2023); Jing et al. (2024); Han et al. (2024); Yu et al. (2024)

Table 12 Comparative overview of adversarial robustness datasets/platforms relevant to autonomous systems, including simulation tools and real-world data

Dataset (References)	Domain	Scenario(s)	Relevance to AS	Use
MNIST Lecun et al. (1998)	Handwritten digits	Baseline testing, digital adversarial examples	Low	Testing classifier vulnerability
CIFAR-10 Krizhevsky (2009)	Small objects, digital images	Digital adversarial attacks, classifier benchmarks	Low	Small-scale adversarial robustness
ImageNet Deng et al. (2009)	Large-scale digital images	Digital adversarial attacks, corruptions	Moderate	Pretraining, digital attack transfer, accuracy drop
ImageNet-P Hendrycks et al. (2021)	Perturbation-augmented ImageNet	Corruptions, robustness evaluation	Moderate	Benchmark for perturbation robustness
COCO, xView Liu et al. (2022)	Object detection	Adversarial patch attacks, digital detection	Moderate	mAP degradation under localized attacks
AD-E20K Zhang et al. (2022c)	Scene segmentation (digital)	Certified patch detection, segmentation	Moderate	Certified accuracy, visual overlap
DOTA Xia et al. (2018)	Aerial images, object detection	Patch attacks, adversarial detection	High	UAV surveillance robustness
Mapillary Traffic Sign Poggi and Mattoccia (2017)	Real-world traffic scenes	Physical adversarial attacks (signs)	High	Traffic sign robustness, AV testing
RobustBench Croce et al. (2020)	Digital, standardized benchmark	Digital adversarial attacks (various datasets)	High	Model benchmarking for adversarial robustness

Table 12 (continued)

Dataset (References)	Domain	Scenario(s)	Relevance to AS	Use
SafeBench Xu et al. (2022)	Simulation (CARLA)	Adversarial scenarios, hostile agents (vehicles/pedestrians)	High	Closed-loop AV safety, collision rate, completion, rule violation
CARLA-GeAR Nesti et al. (2022)	Simulation (CARLA)	Physically-realizable patches on vehicles, adversarial scenarios	High	Multi-task driving (segmentation, detection), mIoU, mAP, depth error
Robust-E2E Jiang et al. (2024)	Simulation (CARLA), E2E driving	White-box input/feature perturbations, corruptions	High	Steering error, lane keeping, success rate under attack
DCI Dataset Zhang et al. (2023)	Simulation + rendering, vehicle detection	Physical patches, weather/angle variations	High	mAP drop, detection under physical attacks
DD-Robust-Bench Wu et al. (2025)	Digital dataset distillation	Digital adversarial attacks, distillation robustness	Moderate	Robustness of distilled datasets
Car Hacking Kang et al. (2021)	Real (CAN logs)	Spoofed/malicious CAN bus messages	High	In-vehicle intrusion detection, false alarm rate
V2X-Sim Li et al. (2023); Zhao et al. (2024b)	Simulation (LiDAR/V2X)	LiDAR spoofing, anomaly injection, cooperative attacks	High	Detection rate, anomaly precision/recall
KITTI-Adv/Blind, STF Lou et al. (2023)	Real+Synth, sensor fusion	Sensor blinding, vision fusion, uncertainty estimation	High	mIoU, mAP under blinding or fusion attacks
DAIR-V2X Zhao et al. (2024b)	Real-world cooperative AV	Malicious contributor, V2X patch attacks	High	Detection accuracy for V2X fusion, anomaly detection
Google Street View Etim and Szefer (2024)	Real images, street scenes	Time-inconsistent, physical perturbations	Moderate–High	Historical adversarial analysis, sign recognition, detection accuracy

Acknowledgements This research is supported, in part, by the UKRI Trustworthy Autonomous Systems Node in Security/EPSRC Grant EP/V026763/1.

Author contributions A.L.P. served as the lead author and was responsible for the conceptualisation, literature review, taxonomy development, methodology, analysis, drafting of the manuscript, visualisation, and overall project administration. P.A. and N.S. reviewed the manuscript critically for important intellectual content; P.A. additionally helped with project administration and editorial communication N.S. additionally assisted in rewriting and sharpening the title and abstract to improve clarity and framing. All authors have read and approved the final manuscript.

References

Abdu-Aguye MG, Gomaa W, Makihara Y, Yagi Y (2020) Detecting adversarial attacks in time-series data. In: ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). https://doi.org/10.1109/icassp40776.2020.9053311

Ai S, Koe ASV, Huang T (2021) Adversarial perturbation in remote sensing image recognition. Appl Soft Comput 105:107252. https://doi.org/10.1016/j.asoc.2021.107252

Akhtar N, Mian A, Kardan N, Shah M (2021) Advances in adversarial attacks and defenses in computer vision: a survey. arXiv preprint arXiv:2108.00401 [cs.CV]

Almutairi S, Barnawi A (2023) Securing dnn for smart vehicles: an overview of adversarial attacks, defenses, and frameworks. J Eng Appl Sci 70(1):1–29. https://doi.org/10.1186/s44147-023-00184-x

Amirkhani A, Karimi MP, Banitalebi-Dehkordi A (2023) A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles. Vis Comput 39:5293–5307. https://doi.org/10.1007/s00371-022-02660-6

Andriushchenko M, Croce F, Flammarion N et al (2020) Square attack: a query-efficient black-box adversarial attack via random search. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds.) Computer vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12368. Springer, Cham. https://doi.org/10.1007/978-3-030-58592-1_29

Angelov P, Soares E (2021) Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xclass). Neural Comput Appl 33:15145–15157. https://doi.org/10.1007/s00521-021-06137-w

Athalye A, Engstrom L, Ilyas A, Kwok K (2017) Synthesizing robust adversarial example. arXiv preprint arXiv:1707.07397 [cs.CV]

Badjie B, Cecílio J, Casimiro A (2024) Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review. ACM Comput Surv 57(1):20. https://doi.org/10.1145/3691625

Bansal M, Krizhevsky A, Ogale A (2018) Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. arXiv preprint arXiv:1812.03079 [cs.RO]

Bekey GA (2005) Autonomous robots: from biological inspiration to implementation and control (intelligent robotics and autonomous agents). The MIT Press, Cambridge. https://doi.org/10.5555/1088950

Besl PJ (1988) Active optical range imaging sensors. Mach Vis Appl 1(2):127–152. https://doi.org/10.1007/BF01212277

Bhagoji AN, He W, Li B, Song D (2017) Exploring the space of black-box attacks on deep neural networks. arXiv preprint arXiv:1712.09491 [cs.LG]

Biggio B, Nelson BA, Laskov P (2012) Poisoning attacks against support vector machines. In: Proceedings of the 29th international conference on international conference on machine learning. ICML'12, pp. 1467–1474. Omni press, Madison. https://doi.org/10.5555/3042573.3042761

Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 [cs.CV]

Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J, Zieba K (2016) End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 [cs.CV]

Bojarski M, Yeres P, Choromanska A, Choromanski K, Firner B, Jackel L, Muller U (2017) Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 [cs.CV]

Boltachev E (2024) Potential cyber threats of adversarial attacks on autonomous driving models. J Comput Virol Hack Tech 20:363–373. https://doi.org/10.1007/s11416-023-00486-x

Brendel W, Rauber J, Bethge M (2017) Decision-based adversarial attacks: reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 [stat.ML]

Brown TB, Mané D, Roy A, Abadi M, Gilmer J (2018) Adversarial patch. arXiv preprint arXiv:1712.09665 [cs.CV]

Bui AT, Le T, Zhao H, Montague P, DeVel O, Abraham T, Phung D (2021) Improving ensemble robustness by collaboratively promoting and demoting adversarial robustness. Proc AAAI Conf Artif Intell 35(8):6831–6839. https://doi.org/10.1609/aaai.v35i8.16843

Cao Y, Xiao C, Cyr B, Zhou Y, Park W, Rampazzi S, Chen QA, Fu K, Mao ZM (2019) Adversarial sensor attack on lidar-based perception in autonomous driving. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (CCS '19). ACM, London, pp. 2267–2281.https://doi.org/10.1145/3319535.3339815

Cao Y, Wang N, Xiao C, Yang D, Fang J, Yang R, Chen QA, Liu M, Li B (2020) Demonstration: 3d adversarial object against msf-based perception in autonomous driving. In: Proceedings of the 3rd conference on machine learning and systems (MLSys). https://me.ningfei.org/paper/MLsys_demo.pdf

Cao Y, Xiao C, Anandkumar A et al (2022) Advdo: Realistic adversarial attacks for trajectory prediction. In: Computer vision – ECCV 2022. Springer, Cham, pp 36–52. https://doi.org/10.1007/978-3-031-20065-6_3

Cao J, Zhang K-Y, Yao T, Ding S, Yang X, Ma C (2024) Towards unified defense for face forgery and spoofing attacks via dual space reconstruction learning. Int J Comput Vis. https://doi.org/10.1007/s11263-024-02151-2

Carlini N, Wagner D (2017a) Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. AISec '17. Association for Computing Machinery, New York, pp. 3–14. https://doi.org/10.1145/3128572.3140444

Carlini N, Wagner D (2017b) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (SP), pp. 39–57. https://doi.org/10.1109/SP.2017.49

Chahe A, Wang C, Jeyapratap A, Xu K, Zhou L (2023) Dynamic adversarial attacks on autonomous driving systems. arXiv preprint arXiv:2312.06701 [cs.RO]

Chen Y, Chu S (2023) Adversarial defense in aerial detection. In: CVPR workshop on adversarial ML. https://doi.org/10.1109/cvprw59228.2023.00226

Chen E-C, Lee C-R (2021) Towards fast and robust adversarial training for image classification. In: Computer vision – ACCV 2020. Springer, Cham. https://doi.org/10.1007/978-3-030-69535-4_35

Chen P-Y, Zhang H, Sharma Y, Yi J, Hsieh C-J (2017) Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. AISec '17. Association for Computing Machinery, New York, pp. 15–26.https://doi.org/10.1145/3128572.3140448

Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018a) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Computer vision – ECCV 2018. Springer, Cham, pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49

Chen P-Y, Sharma Y, Zhang H, Yi J, Hsieh C-J (2018b) Ead: Elastic-net attacks to deep neural networks via adversarial examples. Proc AAAI Conf Artif Intell. https://doi.org/10.1609/aaai.v32i1.11302

Chen S-T, Cornelius C, Martin J, Chau DH (2019) Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. arXiv preprint arXiv:1804.05810 [cs.CV]

Chen J, Jordan MI, Wainwright MJ (2020) Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 IEEE symposium on security and privacy (SP), pp. 1277–1294. https://doi.org/10.1109/SP40000.2020.00045

Chen X, Huang W, Guo W, Zhang F, Du J, Zhou Z (2024a) Adversarial defence by learning differentiated feature representation in deep ensemble. Mach Vis Appl 35(1):88. https://doi.org/10.1007/s00138-024-01571-x

Chen H, Yan H, Yang X, Su H, Zhao S, Qian F (2024b) Efficient adversarial attack strategy against 3d object detection in autonomous driving. IEEE Trans Intell Transp Syst 25(11):16118–16132. https://doi.org/10.1109/TITS.2024.3410038

Cheng S, Liu Y, Ma S, Zhang X (2021) Deep feature space trojan attack of neural networks by controlled detoxification. Proce AAAI Conf Artif Intell 35(2):1148–1156. https://doi.org/10.1609/aaai.v35i2.16201

Codevilla F, Müller M, López A et al (2018) End-to-end driving via conditional imitation learning. In: Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA). IEEE, Brisbane, pp. 4693–4700. https://doi.org/10.1109/ICRA.2018.8460487

Cohen JM, Rosenfeld E, Kolter JZ (2019) Certified adversarial robustness via randomized smoothing. arXiv preprint arXiv:1902.02918 [cs.LG]

Costa JC, Roxo T, Proença H et al (2024) How deep learning sees the world: a survey on adversarial attacks & defenses. IEEE Access 12:61113–61136. https://doi.org/10.1109/ACCESS.2024.3395118

Croce F, Hein M (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the 37th international conference on machine learning. JMLR.org, Virtual Event, p. 206. https://doi.org/10.5555/3524938.3525144

Croce F, Andriushchenko M, Sehwag V, Debenedetti E, Flammarion N, Chiang M, Mittal P, Hein M (2020) Robustbench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670 [cs.LG]

Dash P, Chan E, Pattabiraman K (2024) Specguard: Specification aware recovery for robotic autonomous vehicles from physical attacks. In: Proceedings of the ACM conference on computer and communications security (CCS). https://doi.org/10.1145/3658644.3690210

Deng Y, Mu T (2023) Understanding and improving ensemble adversarial defense. In: Proceedings of the 37th international conference on neural information processing systems. NIPS '23. Curran Associates Inc., Red Hook, NY, USA. https://doi.org/10.5555/3666122.3668653

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deng Y, Zhang T, Lou G, Zheng X, Jin J, Han Q-L (2021) Deep learning-based autonomous driving systems: a survey of attacks and defenses. IEEE Trans Ind Inf 17(12):7897–7912. https://doi.org/10.1109/TII.2021.3071405

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 [cs.CV]

Drucker H, Cun YL (1992) Improving generalization performance using double backpropagation. IEEE Trans Neural Netw 3(6):991–997. https://doi.org/10.1109/72.165600

Du X, Pun C-M, Zhang Z (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Proceedings of the 32nd international conference on neural information processing systems. NIPS'18. Curran Associates Inc., Red Hook, pp. 7167–7177. https://doi.org/10.5555/3327757.3327819

Du A, Chen B, Chin T-J, Law YW, Sasdelli M, Rajasegaran R, Campbell D (2022) Physical adversarial attacks on an aerial imagery object detector. In: 2022 IEEE/CVF winter conference on applications of computer vision (WACV), pp. 3798–3808. https://doi.org/10.1109/WACV51458.2022.00385

Dursun HE, Güven Y, Kumbasar T (2025) Imitation learning for autonomous driving: insights from real-world testing. arXiv preprint, arXiv:2504.18847 [cs.RO]

Edelkamp S (2023) Adversarial planning. Springer, Cham, pp. 325–335. https://doi.org/10.1007/978-3-319-65596-3_18

Ehlers R (2017) Formal verification of piece-wise linear feed-forward neural networks. arXiv preprint, arXiv:1705.01320 [cs.LO]

Engstrom L, Tran B, Tsipras D, Schmidt L, Madry A (2019) Exploring the landscape of spatial robustness. arXiv preprint, arXiv:1712.02779 [cs.LG]

Etim A, Szefer J (2024) Time traveling to defend against adversarial example attacks in image classification. arXiv preprint arXiv:2410.08338 [cs.CR]

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, Prakash A, Kohno T, Song D (2018) Robust physical-world attacks on deep learning visual classification. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp. 1625–1634.https://doi.org/10.1109/CVPR.2018.00175

Feinman R, Curtin RR, Shintre S, Gardner AB (2017) Detecting adversarial samples from artifacts. arXiv preprint, arXiv:1703.00410 [stat.ML]

Fletcher R (2013) Practical methods of optimization, 2nd edn. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9781118723203

Forsyth DA, Ponce J (2011) Computer vision: a modern approach, 2nd edn. Pearson, Boston. https://www.pearson.com/store/p/computer-vision-a-modern-approach/P100000687361/9780136085928

Freitas S, Chen S-T, Wang ZJ, Chau DH (2020) Unmask: Adversarial detection and defense through robust feature alignment. In: 2020 IEEE international conference on big data (Big Data). https://doi.org/10.1109/bigdata50022.2020.9378303

Fu C, Li S, Yuan X, Ye J, Cao Z, Ding F (2022) Ad2attack: Adaptive adversarial attack on real-time uav tracking. In: 2022 International conference on robotics and automation (ICRA), pp. 5893–5899. https://doi.org/10.1109/ICRA46639.2022.9812056

Girdhar M, Hong J, Moore J (2023) Cybersecurity of autonomous vehicles: a systematic literature review of adversarial attacks and defense models. IEEE Open J Vehic Tech PP:1–23. https://doi.org/10.1109/OJVT.2023.3265363

Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Goldblum M, Fowl L, Feizi S, Goldstein T (2020) Adversarially robust distillation. Proc AAAI Conf Artif Intell 34(04):3996–4003. https://doi.org/10.1609/aaai.v34i04.5816

Gong Y, Wang S, Jiang X, Yin L, Sun F (2023) Adversarial example detection using semantic graph matching. Appl Soft Comput 141:110317. https://doi.org/10.1016/j.asoc.2023.110317

Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 [stat.ML]

Gowal S, Dvijotham K, Stanforth R, Bunel R, Qin C, Uesato J, Arandjelovic R, Mann T, Kohli P (2018) On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715 [cs.LG]

Grosse K, Manoharan P, Papernot N, Backes M, McDaniel P (2017) On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 [cs.CR]
Gu T, Dolan-Gavitt B, Garg S (2017) Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 [cs.CR]
Gu T, Liu K, Dolan-Gavitt B, Garg S (2019) Badnets: evaluating backdooring attacks on deep neural networks. IEEE Access 7:47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068
Guesmi A, Hanif MA, Shafique M (2023) Advrain: adversarial raindrops to attack camera-based smart vision systems. Information 14(12):634. https://doi.org/10.3390/info14120634
Guizzo E (2011) How Google's self-driving car works. IEEE spectrum. Accessed: 2025-05-23. https://spectrum.ieee.org/star-autonomous-surgical-robot
Guo F, Zhao Q, Li X, Kuang X, Zhang J, Han Y, Tan Y-A (2019a) Detecting adversarial examples via prediction difference for deep neural networks. Inf Sci 501:182–192. https://doi.org/10.1016/j.ins.2019.05.084
Guo C, Gardner JR, You Y, Wilson AG, Weinberger KQ (2019b) Simple black-box adversarial attacks. arXiv preprint arXiv:1905.07121 [cs.LG]
Gupta KD, Dasgupta D, Akhtar Z (2020) Adversarial input detection using image processing techniques (ipt). In: 2020 IEEE annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE, Virtual Conference, pp. 309–315. https://doi.org/10.1109/UEMCON51285.2020.9298060
Hanfeld P, Höhne MM-C, Bussmann M et al (2023) Flying adversarial patches: manipulating the behavior of deep learning-based autonomous multirotors. arXiv preprint arXiv:2305.12859 [cs.RO]
Han X, Wang H, Zhao K, Deng G, Xu Y, Liu H, Qiu H, Zhang T (2024) Visionguard: Secure and robust visual perception of autonomous vehicles in practice. In: CCS. https://doi.org/10.1145/3658644.3670296
Hao C, Orlando D, Liu J, Yin C (2002) Introduction to radar systems, 3rd edn. McGraw-Hill Education, New York, NY, USA. https://doi.org/10.1007/978-981-16-6399-4_1
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
He K, Gkioxari G, Dollár P et al (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322
Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 [cs.LG]
Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, Desai R, Zhu T, Parajuli S, Guo M, Song D, Steinhardt J, Gilmer J (2021) The many faces of robustness: a critical analysis of out-of-distribution generalization. In: 2021 IEEE/CVF international conference on computer vision (ICCV). https://doi.org/10.1109/iccv48922.2021.00823
Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 [stat.ML]
Hong H, Zhang X, Wang B, Ba Z, Hong Y (2024) Certifiable black-box attacks with randomized adversarial examples: breaking defenses with provable confidence. In: Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security, pp. 600–614. https://doi.org/10.1145/3658644.3690343
Horton E, Ranganathan P (2018) Development of a gps spoofing apparatus to attack a dji matrice 100 quadcopter. J Global Position Syst 16:9. https://doi.org/10.1186/s41445-018-0018-3
Hsiao T-F, Huang B-L, Ni Z-X, Lin Y-T, Shuai H-H, Li Y-H, Cheng W-H (2024) Natural light can also be dangerous: traffic sign misinterpretation under adversarial natural light attacks. In: 2024 IEEE/CVF winter conference on applications of computer vision (WACV), pp. 3903–3912. https://doi.org/10.1109/WACV57701.2024.00387
Hsu J-M (2002) Introduction to global satellite positioning system (GPS). Artech House, Boston, MA. https://doi.org/10.4018/978-1-60566-840-6.ch007
Huang WR, Geiping J, Fowl L et al (2020) Metapoison: Practical general-purpose clean-label data poisoning. In: Proceedings of the 34th international conference on neural information processing systems. NIPS'20. Curran Associates Inc., Red Hook, p. 1013. https://doi.org/10.5555/3495724.3496737
Huang B, Ke Z, Wang Y, Wang W, Shen L, Liu F (2021) Adversarial defence by diversified simultaneous training of deep ensembles. Proc AAAI Conf Artif Intell 35(9):7823–7831. https://doi.org/10.1609/aaai.v35i9.16955
Hu Z, Chu W, Zhu X, Zhang H, Zhang B, Hu X (2023a) Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. arXiv preprint arXiv:2307.01778 [cs.CV]
Hu X, Ying Q, Qian Z, Li S, Zhang X (2023b) Draw: Defending camera-shooted raw against image manipulation. In: 2023 IEEE/CVF international conference on computer vision (ICCV), pp. 22377–22387. https://doi.org/10.1109/iccv51070.2023.02050

Hussain M, Hong J-E (2023) Reconstruction-based adversarial attack detection in vision-based autonomous driving systems. Mach Learn Knowl Extract 5(4):1589–1611. https://doi.org/10.3390/make5040080
Ibrahum ADM, Hussain M, Hong J-E (2024) Deep learning adversarial attacks and defenses in autonomous vehicles: a systematic literature review from a safety perspective. Artif Intell Rev 58:28. https://doi.org/10.1007/s10462-024-11014-8
Jallepalli D, Ravikumar NC, Badarinath PV, Uchil S, Suresh MA (2021) Federated learning for object detection in autonomous vehicles. In: 2021 IEEE seventh international conference on big data computing service and applications (BigDataService), pp. 107–114. https://doi.org/10.1109/BigDataService52369.2021.00018
Janai J, Güney F, Behl A, Geiger A (2020) Computer vision for autonomous vehicles: problems, datasets and state of the art. Found Trends Comput Graph Vis 12(1–3):1–308. https://doi.org/10.1561/0600000079
Jiang W, Wang L, Zhang T, Chen Y, Dong J, Bao W, Zhang Z, Fu Q (2024) Robuste2e: exploring the robustness of end-to-end autonomous driving. Electronics 13(16):3299. https://doi.org/10.3390/electronics13163299
Jing L, Wang R, Ren W, Dong X, Zou C (2024) Pad: Patch-agnostic defense against adversarial patch attacks. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 24472–24481. https://doi.org/10.1109/cvpr52733.2024.02310
Kang H, Kwak BI, Lee YH, Lee H, Lee H, Kim HK (2021) Car hacking: attack & defense challenge 2020 dataset. https://doi.org/10.21227/qvr7-n418
Kannan H, Kurakin A, Goodfellow I (2018) Adversarial logit pairing. arXiv preprint arXiv:1803.06373 [cs.LG]
Katz G, Barrett C, Dill DL et al (2017) Reluplex: An efficient smt solver for verifying deep neural networks. In: Computer aided verification. CAV 2017. Lecture Notes in Computer Science, vol. 10426. Springer, cham. https://doi.org/10.1007/978-3-319-63387-9
Khamaiseh SY, Bagagem D, Al-Alaj A, Mancino M, Alomari HW (2022) Adversarial deep learning: a survey on adversarial attacks and defense mechanisms on image classification. IEEE Access 10:102266–102291. https://doi.org/10.1109/ACCESS.2022.3208131
Khan IA, Moustafa N, Pi D, Haider W, Li B, Jolfaei A (2022) An enhanced multi-stage deep learning framework for detecting malicious activities from autonomous vehicles. IEEE Trans Intell Transp Syst 23(12):25469–25478. https://doi.org/10.1109/TITS.2021.3105834
King I, Wang J (2019) Provably robust deep learning via adversarially trained smoothed classifiers. In: Proceedings of the 33rd international conference on neural information processing systems. Curran Associates Inc., Red Hook, p. 1013. https://doi.org/10.5555/3454287.3455300
Kinsler LE, Frey AR, Coppens AB et al (2000) Fundamentals of acoustics, 4th edn. John Wiley & Sons, Wiley Online Library. https://doi.org/10.1007/978-3-540-48830-9_2
Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick R (2023) Segment anything. In: 2023 IEEE/CVF international conference on computer vision (ICCV). https://doi.org/10.1109/iccv51070.2023.00371
Knee P (2005) Radar signal processing fundamentals. McGraw-Hill, New York, NY, USA. https://doi.org/10.1007/978-3-031-01519-9_4
Koenig S, Bonet B, Cavazza M, desJardins M, Felner A, Hawes N, Knox B, Konidaris G, Lang J, López CL, Magazzeni D, McGovern A, Natarajan S, Sturtevant NR, Thielscher M, Yeoh W, Sardina S, Wagstaff K (2015) Using machine teaching to identify optimal training-set attacks on machine learners. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. AAAI'15. AAAI Press, Austin, pp. 2871–2877. https://doi.org/10.5555/2886521.2886721
Komkov S, Petiushko A (2021) Advhat: Real-world adversarial attack on arcface face id system. In: 2020 25th International conference on pattern recognition (ICPR), pp. 819–826.https://doi.org/10.1109/icpr48806.2021.9412236
Kong Z, Guo J, Li A, Liu C (2020) Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR).https://doi.org/10.1109/cvpr42600.2020.01426
Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical report, University of Toronto. https://www.cs.toronto.edu/∼kriz/learning-features-2009-TR.pdf
Kurakin A, Goodfellow I, Bengio S (2016) Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 [cs.CV]
Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S (2019) Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE symposium on security and privacy (SP), pp. 656–672.https://doi.org/10.1109/SP.2019.00044

Lempitsky V, Vedaldi A, Ulyanov D (2018) Deep image prior. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp. 9446–9454. https://doi.org/10.1109/CVPR.2018.00984
Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J (2018) Defense against adversarial attacks using high-level representation guided denoiser. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp. 1778–1787. https://doi.org/10.1109/CVPR.2018.00191
Li X, Li J, Chen Y, Ye S, He Y, Wang S, Su H, Xue H (2021) Qair: Practical query-efficient black-box attacks for image retrieval. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 3329–3338. https://doi.org/10.1109/CVPR46437.2021.00334
Li Z, Li H, Xie E, Sima C, Lu T, Qiao Y, Dai J (2022) Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Computer vision – ECCV 2022. Springer, Cham, pp. 1–18. https://doi.org/10.1007/978-3-031-20077-9_1
Li Y, Fang Q, Bai J, Chen S, Juefei-Xu F, Feng C (2023) Among us: Adversarially robust collaborative perception by consensus. In: IEEE international conference on computer vision (ICCV). https://doi.org/10.1109/iccv51070.2023.00024
Li L, Qiu J, Spratling M (2024a) Aroid: improving adversarial robustness through online instance-wise data augmentation. Int J Comput Vis 132:1–20. https://doi.org/10.1007/s11263-024-02206-4
Li Y, Angelov P, Suri N (2024b) Self-supervised representation learning for adversarial attack detection. In: Computer vision – ECCV 2024, pp. 236–252. https://doi.org/10.1007/978-3-031-73027-6_14
Li Y, Angelov P, Yu Z, Pellicer AL, Suri N (2024c) Federated adversarial learning for robust autonomous landing runway detection. In: Artificial neural networks and machine learning – ICANN 2024. Lecture Notes in Computer Science. Springer, Cham, vol. 15021, pp. 159–173.https://doi.org/10.1007/978-3-031-72347-6_11
Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 [cs.LG]
Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016a) Ssd: Single shot multibox detector. In: Computer vision – ECCV 2016. Springer, Cham, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
Liu Y, Chen X, Liu C, Song D (2016b) Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 [cs.LG]
Liu N, Du M, Guo R, Liu H, Hu X (2021) Adversarial attacks and defenses: an interpretation perspective. SIGKDD Explor Newsl 23(1):86–99. https://doi.org/10.1145/3468507.3468519
Liu J, Levine A, Lau CP, Chellappa R, Feizi S (2022) Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection. In: CVPR, pp. 14973–14982. https://doi.org/10.1109/cvpr52688.2022.01455
Lou Y, Song Q, Xu Q, Tan R, Wang J (2023) Uncertainty-encoded multi-modal fusion for robust object detection in autonomous driving. In: Proc. of 26th European conference on artificial intelligence (ECAI). https://doi.org/10.3233/faia230441
Lu Z, Sun H, Ji K, Kuang G (2023) Adversarial robust aerial image recognition based on reactive-proactive defense framework with deep ensembles. Remote Sens 15(19):4660. https://doi.org/10.3390/rs15194660
Lu Z, Sun H, Xu Y (2023) Adversarial robustness enhancement of uav-oriented automatic image recognition via ensemble defense. Remote Sens 15(12):3007. https://doi.org/10.3390/rs15123007
Lu X, Radha H (2023) ScAR: scaling adversarial robustness for LiDAR object detection. In: Proc. of IROS. https://doi.org/10.1109/iros55552.2023.10341583
Lu J, Sibai H, Fabry E, Forsyth D (2017) No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:1707.03501 [cs.CV]
Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2019) Towards deep learning models resistant to adversarial attacks. arXiv preprint, arXiv:1706.06083 [stat.ML]
Malik J, Muthalagu R, Pawar PM (2024) A systematic review of adversarial machine learning attacks, defensive controls, and technologies. IEEE Access 12:99382–99421. https://doi.org/10.1109/ACCESS.2024.3423323
Mani N, Moh M, Moh T-S (2019) Towards robust ensemble defense against adversarial examples attack. In: 2019 IEEE global communications conference (GLOBECOM). IEEE, Hawaii, pp. 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9013408
Man Y, Li M, Gerdes R (2020) Ghostimage: Remote perception attacks against camera-based image classification systems. In: 23rd International symposium on research in attacks, intrusions and defenses (RAID 2020), pp. 317–332. USENIX Association, Virtual Conference. https://www.usenix.org/system/files/raid20-man.pdf

Man Y, Li M, Gerdes R (2023) Remote perception attacks against camera-based object recognition systems. In: Proceedings of the 2023 ACM SIGSAC conference on computer and communications security (CCS '23), pp. 14–11422. https://doi.org/10.1145/3596221

Ma S, Vemprala S, Wang W, Gupta JK, Song Y, McDuff D, Kapoor A (2022) Compass: Contrastive multimodal pretraining for autonomous systems. arXiv preprint, arXiv:2203.15788 [cs.RO]

Meng D, Chen H (2017) Magnet: A two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. CCS '17. Association for Computing Machinery, New York, pp. 135–147. https://doi.org/10.1145/3133956.3134057

Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 2574–2582. https://doi.org/10.1109/CVPR.2016.282

Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 86–94. https://doi.org/10.1109/CVPR.2017.17

Morgulis N, Kreines A, Mendelowitz S, Weisglass Y (2019) Fooling a real car with adversarial traffic signs. arXiv preprint arXiv:1907.00374 [cs.CR]

Mu J (2024) A real-time defense against object vanishing adversarial patch attacks for object detection in autonomous vehicles. arXiv preprint, arXiv:2412.06215 [cs.CV]

Muravev N, Petiushko A (2022) Certified robustness via randomized smoothing over multiplicative parameters of input transformations. In: Proceedings of the thirty-first international joint conference on artificial intelligence, pp. 3366–3372. https://doi.org/10.24963/ijcai.2022/467

Naseer M, Khan SH, Porikli F (2018) Local gradients smoothing: defense against localized adversarial attacks. arXiv preprint arXiv:1807.01216 [cs.CV]

Nesti F, Rossolini G, D'Amico G, Biondi A, Buttazzo G (2022) Carla-gear: a dataset generator for a systematic evaluation of adversarial robustness of vision models. arXiv preprint arXiv:2206.04365 [cs.CV]

Nie W, Guo B, Huang Y, Xiao C, Vahdat A, Anandkumar A (2022) Diffusion models for adversarial purification

Noz-González LM, Pfitzner B, Russo M, Carnerero-Cano J, Lupu EC (2019) Poisoning attacks with generative adversarial nets. arXiv preprint arXiv:1906.07773 [cs.LG]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang P-Y, Li S-W, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P (2024) Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 [cs.CV]

Oslund S, Washington C, So A, Chen T, Ji H (2022) Multiview robust adversarial stickers for arbitrary objects in the physical world. J Comput Cognit Eng 1(4):152–158. https://doi.org/10.47852/bonviewJCCE2202322

Pan Y, Cheng C-A, Saigol K, Lee K, Yan X, Theodorou E, Boots B (2017) Agile autonomous driving using end-to-end deep imitation learning. arXiv preprint arXiv:1709.07174 [cs.RO]

Papernot N, McDaniel P, Goodfellow I (2016a) Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 [cs.CR]

Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016b) The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS &P), pp. 372–387. https://doi.org/10.1109/EuroSP.2016.36

Papernot N, McDaniel P, Wu X, Jha S, Swami A (2016c) Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP), pp. 582–597. https://doi.org/10.1109/SP.2016.41

Pellicer AL, Li Y, Angelov P (2024a) PUDD: Towards robust multi-modal prototype-based deepfake detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops, pp. 3809–3817. https://doi.org/10.1109/CVPRW63382.2024.00385

Pellicer AL, Giatgong K, Li Y, Suri N, Angelov P (2024b) Unicad: A unified approach for attack detection, noise reduction and novel class identification. In: 2024 International joint conference on neural networks (IJCNN), pp. 1–8. https://doi.org/10.1109/ijcnn60899.2024.10651159

Poggi M, Mattoccia S (2017) Learning to predict stereo reliability enforcing local consistency of confidence maps. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 4541–4550. https://doi.org/10.1109/CVPR.2017.483

Pomerleau DA (1988) Alvinn: An autonomous land vehicle in a neural network. In: Proceedings of the 2nd international conference on neural information processing systems, pp. 305–313. https://doi.org/10.5555/2969735.2969771

Pourkeshavarz M, Sabokrou M, Rasouli A (2024) Adversarial backdoor attack by naturalistic data poisoning on trajectory prediction in autonomous driving. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 14885–14894. https://doi.org/10.1109/CVPR52733.2024.01410

Prevot T, Rios J, Kopardekar P, III JER, Johnson M, Jung J (2016) Uas traffic management (utm) concept of operations to safely enable low altitude flight operations. In: 16th AIAA aviation technology, integration, and operations conference. https://doi.org/10.2514/6.2016-3292

Queyrut S, Schiavoni V, Felber P (2023) Mitigating adversarial attacks in federated learning with trusted execution environments. arXiv preprint arXiv:2309.07197 [cs.LG]

Raghunathan A, Steinhardt J, Liang P (2018) Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344 [cs.LG]

Ren Shaoqing, He et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 29th international conference on neural information processing systems - Volume 1, pp. 91–99. https://doi.org/10.5555/2969239.2969250

Renz K, Chen L, Marcu A-M, Hünermann J, Hanotte B, Karnsund A, Shotton J, Arani E, Sinavski O (2024) Carllava: Vision language models for camera-only closed-loop driving. arXiv preprint arXiv:2406.10165 [cs.CV]

Reyes-Amezcua I, Ochoa-Ruiz G, Mendez-Vazquez A (2024) Enhancing image classification robustness through adversarial sampling with delta data augmentation (dda). In: 2024 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp. 274–283. https://doi.org/10.1109/CVPRW63382.2024.00032

Rony J, Hafemann LG, Oliveira LS, Ayed IB, Sabourin R, Granger E (2019) Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4317–4325. https://doi.org/10.1109/CVPR.2019.00445

Ross A, Doshi-Velez F (2018) Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32. https://doi.org/10.1609/aaai.v32i1.11504

Rozsa A, Rudd EM, Boult TE (2016) Adversarial diversity and hard positive generation. In: 2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp. 410–417. https://doi.org/10.1109/CVPRW.2016.58

Sabokrou M, Khalooei M, Adeli E (2024) Be your own neighborhood: detecting adversarial examples by the neighborhood relations built on self-supervised learning. In: Proceedings of the 41st international conference on machine learning (ICML 2024). JMLR.org, Vienna, p. 2794. https://doi.org/10.5555/3692070.3692794

Samangouei P, Kabkab M, Chellappa R (2018) Defense-gan: protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605 [cs.CV]

Shah A (2023) Adversary ml resilience in autonomous driving through human-centered perception mechanisms. arXiv preprint arXiv:2311.01478 [cs.CV]

Shan S, Ding W, Passananti J, Wu S, Zheng H, Zhao BY (2024) Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In: 2024 IEEE symposium on security and privacy (SP), pp. 807–825. https://doi.org/10.1109/sp54263.2024.00207

Sharif M, Bhagavatula S, Bauer L, Reiter MK (2016) Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. CCS '16. Association for Computing Machinery, New York, pp. 1528–1540. https://doi.org/10.1145/2976749.2978392

Sharif M, Bhagavatula S, Bauer L, Reiter MK (2019) A general framework for adversarial examples with objectives. ACM Trans Priv Sec 1(1):1–30. https://doi.org/10.1145/3317611

Shen Y, Zheng L, Shu M et al (2021) Gradient-free adversarial training against image corruption for autonomous driving. In: Proceedings of the 35th international conference on neural information processing systems. Curran Associates Inc., Red Hook, pp. 26250–26263. https://doi.org/10.5555/3540261.3542271

Sheridan TB (2016) Human-robot interaction: status and challenges. Hum Factors 58(4):525–532. https://doi.org/10.1177/0018720816644364

Shibly KH, Hossain MD, Inoue H, Taenaka Y, Kadobayashi Y (2023) Towards autonomous driving model resistant to adversarial attack. Appl Artif Intell 37(1):2193461. https://doi.org/10.1080/08839514.2023.2193461

Shi L, Chen Z, Shi Y, Zhao G, Wei L, Tao Y, Gao Y (2022) Data poisoning attacks on federated learning by using adversarial samples. In: 2022 International conference on computer engineering and artificial intelligence (ICCEAI), pp. 158–162. https://doi.org/10.1109/ICCEAI55464.2022.00041

Shu M, Shen Y, Lin MC, Goldstein T (2021) Adversarial differentiable data augmentation for autonomous systems. In: 2021 IEEE international conference on robotics and automation (ICRA). IEEE, Xi'an, pp. 1032–1038. https://doi.org/10.1109/ICRA48506.2021.9561205

Siciliano B, Khatib O (eds) (2016) Springer handbook of robotics, 2nd. edn. Springer, Cham. https://doi.org/10.1007/978-3-319-32552-1

Soares E, Angelov P, Suri N (2022) Similarity-based deep neural network to detect imperceptible adversarial attacks. In: 2022 IEEE symposium series on computational intelligence (SSCI), pp. 1028–1035. https://doi.org/10.1109/SSCI51031.2022.10022016

Song R, Ozmen MO, Kim H et al (2023) Discovering adversarial driving maneuvers against autonomous vehicles. In: 32nd USENIX security symposium (USENIX Security 23). USENIX Association, Anaheim, CA, USA. https://www.usenix.org/system/files/usenixsecurity23-song.pdf

Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput 23(5):828–841. https://doi.org/10.1109/TEVC.2019.2890858

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 [cs.CV]

Szeliski R (2022) Computer vision: algorithms and applications, 2nd edn. Springer, Cham. https://doi.org/10.1007/978-3-030-34372-9

Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 10781–10790. https://doi.org/10.1109/CVPR42600.2020.01079

Tarchoun B, Khalifa AB, Mahjoub MA, Abu-Ghazaleh N, Alouani I (2023) Jedi: Entropy-based localization and removal of adversarial patches. In: CVPR, pp. 4087–4095. https://doi.org/10.1109/cvpr52729.2023.00398

Tesla Inc (2022) Replacing ultrasonic sensors with tesla vision. Accessed: 2025-05-21. https://www.tesla.com/support/transitioning-tesla-vision

Thys S, Ranst WV, Goedemé T (2019) Fooling automated surveillance cameras: adversarial patches to attack person detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp. 49–55. https://doi.org/10.1109/CVPRW.2019.00012

Tian J, Wang B, Guo R, Wang Z, Cao K, Wang X (2022) Adversarial attacks and defenses for deep-learning-based unmanned aerial vehicles. IEEE Internet Things J 9(22):22399–22409. https://doi.org/10.1109/JIOT.2021.3111024

Tian X, Gu J, Li B, Liu Y, Wang Y, Zhao Z, Zhan K, Jia P, Lang X, Zhao H (2024) Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 [cs.CV]

Tjeng V, Xiao K, Tedrake R (2017) Evaluating robustness of neural networks with mixed integer programming. arXiv preprint arXiv:1711.07356 [cs.LG]

Toheed A, Yousaf MH, Rabnawaz JA (2022) Physical adversarial attack scheme on object detectors using 3d adversarial object. In: 2022 2nd international conference on digital futures and transformative technologies (ICoDT2), pp. 1–4. https://doi.org/10.1109/ICoDT255437.2022.9787422

Tramèr F, Boneh D (2019) Adversarial training and robustness for multiple perturbations. In: Proceedings of the 33rd international conference on neural information processing systems. Curran Associates Inc., Red Hook, p. 527. https://doi.org/10.5555/3454287.3454814

Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: attacks and defences. arXiv preprint arXiv:1705.07204 [stat.ML]

Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning. ICML '08. Association for Computing Machinery, New York, pp. 1096–1103. https://doi.org/10.1145/1390156.1390294

Wang X, Cai M, Sohel F, Sang N, Chang Z (2021) Adversarial point cloud perturbations against 3d object detection in autonomous driving systems. Neurocomputing 466:27–36. https://doi.org/10.1016/j.neucom.2021.09.027

Wang C-Y, Bochkovskiy A, Liao H-YM (2023a) Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 7464–7475. https://doi.org/10.1109/cvpr52729.2023.00721

Wang N, Luo Y, Sato T, Xu K, Chen QA (2023b) Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In: 2023 IEEE/CVF international conference on computer vision (ICCV), pp. 4389–4400.https://doi.org/10.1109/iccv51070.2023.00407

Wang G, Zhou C, Wang Y, Chen B, Guo H, Yan Q (2023c) Beyond boundaries: a comprehensive survey of transferable attacks on ai systems. arXiv preprint arXiv:2311.11796 [cs.CR]

Wang Z, Li X, Zhu H et al (2024a) Revisiting adversarial training at scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 24675–24685. https://doi.org/10.1109/CVPR52733.2024.02330

Wang T, Lu F, Zheng Z, Chen G, Jiang C (2024b) Rcdn: Towards robust camera-insensitivity collaborative perception via dynamic feature-based 3d neural modeling. In: Proceedings of the 38th conference on neural information processing systems (NeurIPS). https://proceedings.neurips.cc/paper_files/paper/2024/file/27e5626cabdbb6cd5c56ce4114ff93e4-Paper-Conference.pdf

Wan Z, Shen J, Chuang J, Xia X, Garcia J, Ma J, Chen QA (2022) Too afraid to drive: systematic discovery of semantic dos vulnerability in autonomous driving planning under physical-world attacks. arXiv preprint arXiv:2201.04610 [cs.CR]
Wong E, Rice L, Kolter JZ (2020) Fast is better than free: revisiting adversarial training. arXiv preprint arXiv:2001.03994 [cs.LG]
Wu Z, Lim S-N, Davis LS et al (2020) Making an invisibility cloak: real world adversarial attacks on object detectors. In: Computer vision – ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Springer, Berlin, pp. 1–17. https://doi.org/10.1007/978-3-030-58548-8_1
Wu Y, Du J, Liu P, Lin Y, Xu W, Cheng W (2025) Dd-robustbench: an adversarial robustness benchmark for dataset distillation. IEEE Trans Image Process 34:2052–2066. https://doi.org/10.1109/tip.2025.3553786
Xia G-S, Bai X, Ding J, Zhu Z, Belongie S, Luo J, Datcu M, Pelillo M, Zhang L (2018) Dota: A large-scale dataset for object detection in aerial images. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp. 3974–3983. https://doi.org/10.1109/CVPR.2018.00418
Xiang C, Mahloujifar S, Mittal P (2022) PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In: 31st USENIX security symposium (USENIX Security 22). USENIX Association, Boston, pp. 2065–2082. https://www.usenix.org/conference/usenixsecurity22/presentation/xiang
Xiao C, Zhu J-Y, Li B, He W, Liu M, Song D (2018) Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 [cs.CR]
Xie C, Wang J, Zhang Z, Ren Z, Yuille A (2017a) Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991 [cs.CV]
Xie C, Wang J, Zhang Z, Zhou Y, Xie L, Yuille A (2017b) Adversarial examples for semantic segmentation and object detection. In: 2017 IEEE international conference on computer vision (ICCV), pp. 1378–1387. https://doi.org/10.1109/ICCV.2017.153
Xie C, Wu Y, Maaten L, Yuille AL, He K (2019) Feature denoising for improving adversarial robustness. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 501–509. https://doi.org/10.1109/CVPR.2019.00059
Xing W, Li M, Li M, Han M (2025) Towards robust and secure embodied ai: a survey on vulnerabilities and attacks. arXiv preprint arXiv:2502.13175 [cs.CR]
Xu W, Evans D, Qi Y (2017) Feature squeezing: detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 [cv.CV]
Xu K, Zhang G, Liu S, Fan Q, Sun M, Chen H, Chen P-Y, Wang Y, Lin X (2020) Adversarial t-shirt! evading person detectors in a physical world. In: Computer vision – ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V. Springer, Berlin, pp. 665–681. https://doi.org/10.1007/978-3-030-58558-7
Xu C, Ding W, Lyu W, Liu Z, Wang S, He Y, Hu H, Zhao D, Li B (2022) Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. arXiv preprint arXiv:2206.09682 [cs.RO]
Xu Y, Hu Y, Zhang Z, Meyer GP, Mustikovela SK, Srinivasa S, Wolff EM, Huang X (2024) Vlm-ad: End-to-end autonomous driving through vision-language model supervision. arXiv preprint arXiv:2412.14446 [cs.CV]
Yang H, Zhao J, Xiong Z, Lam K-Y, Sun S, Xiao L (2021) Privacy-preserving federated learning for uav-enabled networks: learning-based joint scheduling and resource management. IEEE J Sel Areas Commun 39(10):3144–3159. https://doi.org/10.1109/jsac.2021.3088655
Yang J, Kim H, Wan W, Hovakimyan N, Vorobeychik Y (2023) Certified robust control under adversarial perturbations. arXiv preprint arXiv:2302.02208
Yeong DJ, Velasco-Hernandez G, Barry J, Walsh J (2021) Sensor and sensor fusion technology in autonomous vehicles: a review. Sensors 21(6):2140. https://doi.org/10.3390/s21062140
Yin J-L, Wang WL, Lin W, Liu X (2025) Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. Proc AAAI Conf Artif Intell 39(9):9508–9516. https://doi.org/10.1609/aaai.v39i9.33030
Yu S, Hirche M, Huang Y, Chen H, Allgöwer F (2021) Model predictive control for autonomous ground vehicles: a review. Auton Intell Syst. https://doi.org/10.1007/s43684-021-00005-z
Yu Z, Li A, Wen R, Chen Y, Zhang N (2024) Physense: Defending physically realizable attacks for autonomous systems via consistency reasoning. In: Proceedings of the ACM conference on computer and communications security (CCS). https://doi.org/10.1145/3658644.3690236
Zhang Y, Zhang Y, Qi J, Bin K, Wen H, Tong X, Zhong P (2022a) Adversarial patch attack on multi-scale object detection for uav remote sensing images. Remote Sens 14(21):5298. https://doi.org/10.3390/rs14215298
Zhang Q, Hu S, Sun J, Chen QA, Mao ZM (2022b) On adversarial robustness of trajectory prediction for autonomous vehicles. In: CVPR, pp. 15159–15168. https://doi.org/10.1109/cvpr52688.2022.01473

Zhang K, Zhou H, Bian H, Zhang W, Yu N (2022c) Certified defense against patch attacks via mask-guided randomized smoothing. In: Proc. ICLR. https://doi.org/10.1007/s11432-021-3457-7

Zhang T, Xiao Y, Zhang X, Li H, Wang L (2023) Benchmarking the physical-world adversarial robustness of vehicle detection. arXiv preprint arXiv:2304.05098 [cs.CV]

Zhang Y, Liu Z, Jia C, Zhu Y, Miao C (2024) An online defense against object-based lidar attacks in autonomous driving. In: Proceedings of the 22nd ACM conference on embedded networked sensor systems (SenSys). https://doi.org/10.1145/3666025.3699345

Zhao S, Ma X, Zheng X, Bailey J, Chen J, Jiang Y-G (2019) Clean-label backdoor attacks. https://doi.org/10.1109/cvpr42600.2020.01445

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2024a) Detrs beat yolos on real-time object detection. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 16965–16974. https://doi.org/10.1109/cvpr52733.2024.01605

Zhao Y, Xiang Z, Yin S, Pang X, Wang Y, Chen S (2024b) Malicious agent detection for robust multi-agent collaborative perception. In: Proc. of IROS. https://doi.org/10.1109/iros58592.2024.10801337

Zhou H, Li W, Kong Z, Guo J, Zhang Y, Yu B, Zhang L, Liu C (2020) Deepbillboard: Systematic physical-world testing of autonomous driving systems. In: Proceedings of the ACM/IEEE 42nd international conference on software engineering. ICSE '20. Association for Computing Machinery, New York, pp. 347–358. https://doi.org/10.1145/3377811.3380422

Zhu X, Liu Y, Hu Z, Li J, Hu X (2024) Infrared adversarial car stickers. arXiv preprint arXiv:2405.09924 [cs.CV]