Review

Not peer-reviewed version

From Vulnerability to Robustness: A Survey of Patch Attacks and Defenses in Computer Vision

Xinyun Liu and Ronghua Xu *

Posted Date: 22 October 2025

doi: 10.20944/preprints202510.1706.v1

Keywords: adversarial patch attacks; adversarial defense; computer vision security; deep neural networks; physical adversarial attacks; trustworthy AI

 doi:10.20944/preprints202510.1706.v1

Article

From Vulnerability to Robustness: A Survey of Patch Attacks and Defenses in Computer Vision

Xinyun Liu and Ronghua Xu *

Department of Applied Computing, Michigan Technological University, Houghton, MI 49931, USA
* Correspondence: ronghuax@mtu.edu

Abstract

Adversarial patch attacks have emerged as a powerful and practical threat to machine learning models in vision-based tasks. Unlike traditional perturbation-based adversarial attacks, which often require imperceptible changes to the entire input, patch attacks introduce localized and visible modifications that can consistently mislead deep neural networks across varying conditions. Their physical realizability makes them particularly concerning for real-world security-critical applications. In response, a growing body of research has proposed diverse defense strategies, including input preprocessing, robust model training, detection-based approaches, and certified defense mechanisms. In this paper, we provide a comprehensive review of patch-based adversarial attacks and corresponding defense techniques. We introduce a new, task-oriented taxonomy that systematically categorizes patch attack methods according to their downstream vision applications (e.g., classification, detection, segmentation) and defense mechanisms based on three major strategies. This unified framework provides an integrated perspective that bridges attack and defense research. Furthermore, we highlight open challenges, such as balancing robustness and model utility, addressing adaptive attackers, and ensuring physical-world resilience. Finally, we outline promising research directions to inspire future work toward building trustworthy and robust vision systems against patch-based adversarial threats.

Keywords: adversarial patch attacks; adversarial defense; computer vision security; deep neural networks; physical adversarial attacks; trustworthy AI

1. Introduction

Deep learning (DL) has advanced at an unprecedented pace and fundamentally transformed many areas of computer vision [1–5]. Convolutional Neural Networks (CNNs), in particular, have driven breakthroughs in face recognition, object detection, and scene understanding, achieving performance that in many benchmark settings rivals and in some cases exceeds human accuracy [6,7]. This dramatic success has encouraged widespread deployment of deep models across safety-critical domains such as autonomous driving, medical imaging, and surveillance, making the reliability and trustworthiness of these systems a matter of practical importance [8].

However, high benchmark performance has not eliminated a growing concern: real-world adoption is increasingly hampered by security and robustness issues. Although models attain impressive average accuracy on curated test sets, they often exhibit brittle behavior under unexpected inputs or malicious manipulation [9–11]. A variety of failure modes contribute to a persistent gap between promising laboratory results and dependable field operation. For example, limited interpretability makes it difficult to understand and diagnose why a model fails, which in turn complicates the development of reliable mitigation strategies [12]. Distribution shifts between training and deployment environments can cause a significant drop in performance, as models often fail to generalize beyond the data they were trained on [13]. Implementation constraints, such as limited computational resources, quantization, or latency requirements, can further degrade model robustness and reliability in real-world systems [14]. Among these factors, adversarial vulnerabilities stand out because they directly

 doi:10.20944/preprints202510.1706.v1

undermine trust: an attacker can intentionally craft inputs that force a model to make dangerous or costly mistakes, eroding confidence in automated decision-making [15].

One prominent class of adversarial threats is adversarial examples: carefully constructed perturbations that cause incorrect predictions while remaining imperceptible or inconspicuous to humans [16]. Research on adversarial attacks has progressed rapidly, starting with additive perturbation methods (e.g., PGD [17] and its iterative variants) that add small, often imperceptible noise across an entire image. Subsequent work broadened the threat model to include more realistic scenarios, such as physically realizable perturbations and attacks that exploit natural scene properties. Parallel to these developments, poisoning attacks shifted attention from test-time manipulations to training-time compromises (e.g., data poisoning and backdoors [18]). Therefore, this requires that defenses must consider the full ML lifecycle, including data collection, model training, and reference.

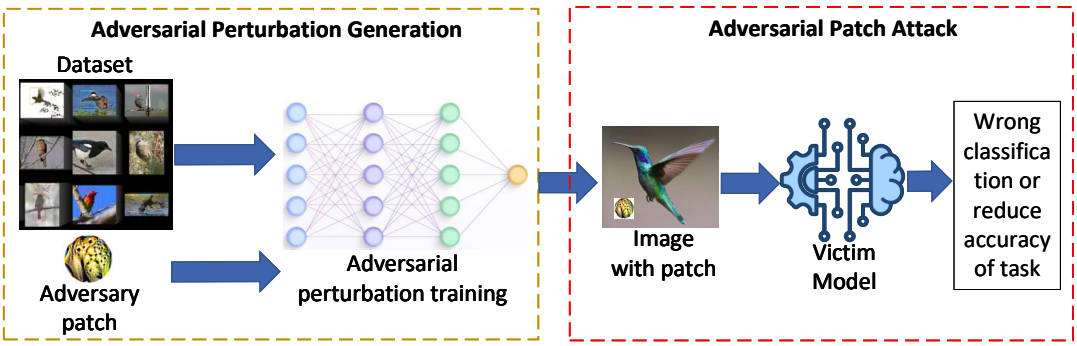

Figure 1. Illustration of the conceptual framework for the adversarial patch attack.

Within this evolving landscape, adversary patch attacks (also called patch-based or localized visible attacks) have emerged as a particularly consequential and practical threat [19]. Figure 1 demonstrates a general framework consisting of adversary perturbation generation and patch attack. Unlike small, distributed noise patterns, a patch attack embeds a localized, contiguous sub-image, often with a distinct pattern or sticker, into the scene. The resulting perturbation is human-visible but strategically designed to trigger misclassification, evade detection, or cause targeted behavior in vision systems. Patch attacks combine several properties that make them especially troubling in real-world contexts: they are physically implementable (a printed sticker suffices), they can operate in black-box settings with limited knowledge of the victim model, and they can be robust to viewpoint changes, lighting, and image transformations [20].

The practical feasibility of patch attacks distinguishes them from many other adversarial strategies. While imperceptible noise attacks are powerful in white-box, digital settings, they are generally fragile when transferred to the physical world. In contrast, patch attacks require only a localized alteration and are easy to manufacture and deploy, which makes them attractive to adversaries with modest resources [21]. The main drawback of patch attacks, their visible nature, is often less of a concern in many real-world settings. In scenarios where human oversight is minimal, such as with remote cameras or autonomous sensors, visibility poses little obstacle. Even in more public environments, attackers can easily disguise patches as ordinary objects like stickers, logos, signs, or pieces of graffiti, making them appear completely natural within the scene [20].

Adversarial patch attacks evolved from earlier, less practical digital attacks that altered entire images imperceptibly, to increasingly sophisticated and physically-realizable attacks for various deep learning systems. Figure 2 summarizes the timeline of adversarial patch attack research from 2017 to 2025. Initial works focused on simple sticker-like patches that misled classifiers or detectors. Subsequent developments incorporated Generative Adversarial Networks (GAN)-based generation for higher fidelity, adaptive optimization to exploit model weaknesses, and stealthy designs minimizing visual footprint. More recent efforts have explored universal and prompt-guided patches, enabling attacks that are both effective and consistent under diverse environmental conditions.

 doi:10.20944/preprints202510.1706.v1

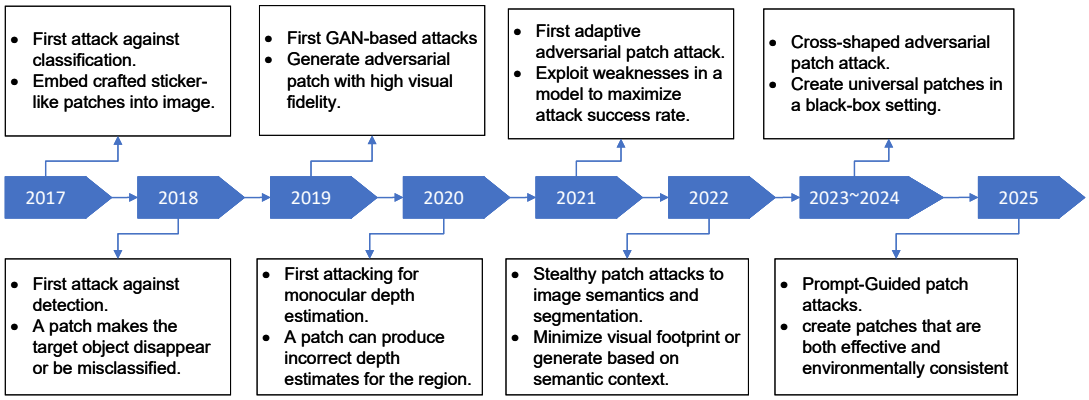

Figure 2. The history of adversary patch attacks.

The arms race between patch attackers and defenders continues. Adversary patch attack methods are becoming more advanced with the use of new technologies like GANs and diffusion models. Given the rising importance of patch attacks, a systematic survey is timely. While a few surveys [20,22,23] have touched upon adversarial machine learning in general, most focus on imperceptible perturbation-based attacks or broad adversarial robustness, often giving limited or outdated coverage of patch-based threats. Furthermore, existing reviews rarely provide a fine-grained taxonomy that captures the unique characteristics of patch attacks, nor do they comprehensively compare defense strategies in terms of practicality, scalability, and physical-world effectiveness. This gap underscores the need for a dedicated, in-depth, and up-to-date survey on patch attacks and defenses in vision-based machine learning.

To this end, this paper aims to provide a timely and comprehensive survey of adversarial patch attacks and defenses for vision tasks. We focus on both digital and physical attacks, their corresponding countermeasures, and evaluation practices. Specifically, we seek to answer the following questions: i) What are the major categories of patch attack methods, and how have they evolved over time? ii) What defense mechanisms have been proposed, and how effective are they across different scenarios?; and iii) What are the key challenges, limitations, and promising future directions for building patch-resilient vision systems?

In this wrok, we review representative papers published between 2015 and 2025. Regarding defense methods, we focus particularly on works from 2023 to 2025. Compared with existing surveys [20,22–24], the main contributions of this paper are as follows:

- Comprehensive and timely review: we present an up-to-date and in-depth survey of adversarial patch attacks and defenses, especially covering works published in the past two years. Compared with existing surveys, our work emphasizes both digital and physical attack settings across a wide range of vision tasks, ensuring timeliness and broader applicability.
- New taxonomy of patch attacks and defenses: we introduce a new, task-oriented taxonomy that systematically categorizes patch attack methods according to their downstream vision applications (e.g., classification, detection, segmentation) and defense mechanisms based on three major strategies. This unified framework provides an integrated perspective that bridges attack and defense research.
- Identification of challenges and future research directions: we summarize open challenges in developing patch-resilient vision systems and discuss promising directions such as adaptive defense frameworks, benchmark standardization, cross-modal robustness, and realistic physical-world evaluations.

The remaining article is organized as follows. Section 2 introduces background of DNN and vision-based tasks and preliminaries of adversary patch attacks. Section 3 presents a taxonomy of patch attacks and reviews representative methods. Section 4 discusses defense mechanisms for patch

 doi:10.20944/preprints202510.1706.v1

attacks. Moreover, we highlights open challenges and future research directions in Section 5. Finally, we conclude this review in Section 6.

2. Background and Related Work

2.1. Deep Neural Networks in Computer Vision

Deep neural networks (DNNs) have become the foundation of modern computer vision, achieving state-of-the-art performance across a wide range of tasks such as image classification, object detection, semantic segmentation, and face recognition [1,2,25]. Convolutional Neural Networks (CNNs) remain the most widely adopted architecture due to their ability to extract hierarchical visual features from local receptive fields, while more recent architectures such as Vision Transformers (ViTs) [5] leverage self-attention mechanisms to capture long-range dependencies and global context. Despite these advances, both CNNs and transformers are inherently vulnerable to adversarial perturbations, including adversarial patches [26]. This vulnerability often stems from their reliance on local discriminative patterns, sensitivity to small but carefully crafted changes, and overemphasis on texture rather than shape. As a result, even visible and localized modifications, such as adversarial patches, can severely compromise their predictions, exposing critical security concerns for vision-based machine learning applications [27].

2.2. Vision-based tasks in ML

Vision-based machine learning tasks aim to enable systems to perceive, interpret, and act upon visual information. Among the most fundamental tasks is image classification, where a single label is assigned to an entire image based on its dominant object or scene [28]. For example, classification models can distinguish between images of cats and dogs, and this task often serves as the foundation for more complex vision applications [29]. Another widely studied task is object detection, which involves both identifying and localizing multiple objects within an image by predicting their categories as well as bounding boxes [30,31]. Typical applications include pedestrian detection in autonomous driving or monitoring in surveillance systems. Also, semantic segmentation provides a pixel-level understanding of an image, assigning each pixel to a semantic category such as road, sky, or vehicle [32]. This fine-grained analysis enables detailed scene interpretation and is critical in domains like robotics and medical imaging. In addition, face recognition represents a specialized yet highly practical task, where the system matches or verifies the identity of individuals based on facial features, often under unconstrained real-world conditions [33,34]. Similar recognition-based tasks extend to other biometric modalities such as gait, iris, and fingerprint analysis. These tasks form the backbone of computer vision research and a wide range of real-world applications. At the same time, their increasing deployment in safety and security-critical domains has made them natural and attractive targets for adversarial patch attacks.

2.3. Adversarial Patch Attacks

Adversarial patches represent a distinct class of adversarial examples. Unlike imperceptible perturbations, which subtly modify pixel values across the entire image, adversarial patches introduce localized, often conspicuous modifications that can be placed onto an image or object, as shown in Figure 1. These patches are typically optimized to maximize the likelihood of misclassification regardless of the surrounding context, making them highly transferable and reusable. Key characteristics of adversarial patches include:

- **Locality:** They affect only a specific region of the input rather than the entire image.
- **Visibility:** The modifications are usually perceptible to humans, yet remain effective against machine learning models.
- **Transferability:** A single patch can generalize across multiple inputs and sometimes across different models.

 doi:10.20944/preprints202510.1706.v1

- **Physical realizability:** Adversarial patches can be printed and deployed in real-world settings, posing threats to autonomous driving, surveillance, and facial recognition.
- **Reusability:** Once generated, the same patch can be applied repeatedly to achieve consistent attack success.

These properties distinguish adversarial patches as one of the most practical and threatening forms of adversarial attacks in vision systems.

Figure 2 demonstrates the history of adversarial patch attacks and highlights their important advances. The first adversarial patch was introduced by Brown et al. [19] in 2017, which caused a DNN classifier to misidentify the object. Following the initial patch for image classifiers, Dpatch [35] either made the object disappear or be misclassified during detection tasks. In 2018, physical patch attach relies on perturbed physical objects that are either ignored or mislabeled by object detection models used by autonomous vehicles [36]. Later, Generative Adversarial Networks (GANs) were used to create more naturalistic-looking patches that were less conspicuous to humans and harder for defense systems to detect [26]. Since 2021, adaptive and stealthy patches have explored advanced attack methods to evade defense systems by minimizing the visual footprint of patches [37] or generating patches based on the semantic context of the scene [38]. In 2023, a cross-shaped adversarial patch attack adopted two line segments to improve patch which has a globalized perturbation capacity while preserving its continuousness [39]. HardBeat [40] can create universal patches in a black-box setting by observing only the predicted labels. Prompt-Guided Patches [41] was proposed in 2025, which uses diffusion models to create patches that are both effective and environmentally consistent, blending in with the background to appear more natural.

To systematically organize the extensive research on adversarial patch attacks, we propose a taxonomy structured along four primary dimensions: attacker knowledge, attack goal, patch properties, and application scenario. This framework, illustrated in Figure 3, provides a comprehensive lens through which to categorize and analyze the vast body of literature.

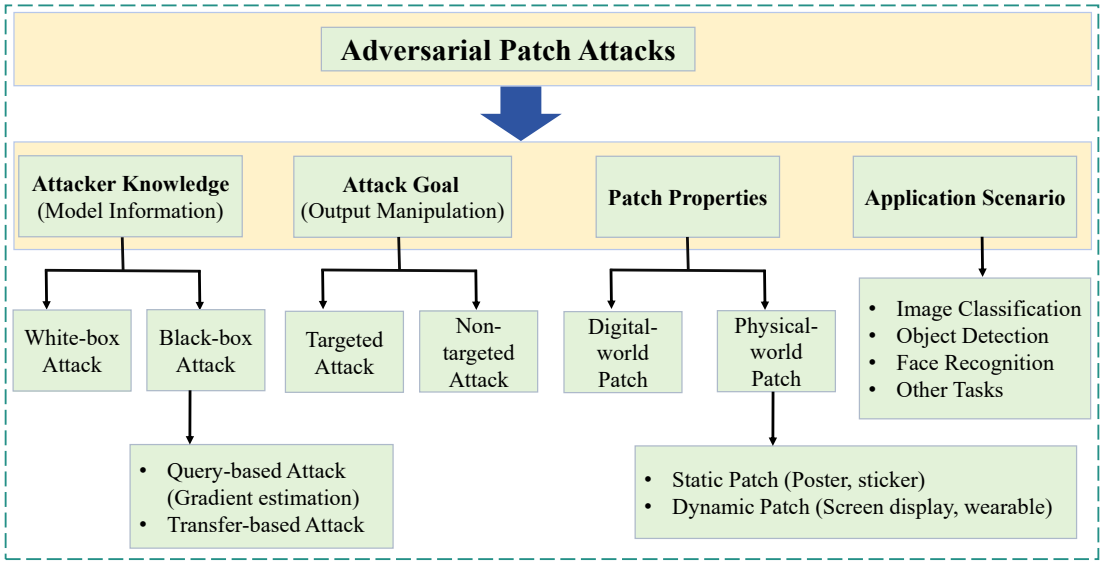

Figure 3. A taxonomy of adversarial patch attacks organized along four key dimensions: attacker knowledge, attack goal, patch properties, and application scenario.

2.3.1. Attacker Knowledge

The first dimension classifies attacks based on the attacker's assumed knowledge of the target model, a critical factor influencing the attack's feasibility and methodology. In a white-box attack setting, the adversary possesses complete knowledge of the model's architecture and parameters [42]. This access allows for the direct computation of gradients via backpropagation, enabling highly

 doi:10.20944/preprints202510.1706.v1

efficient and potent optimization of the patch to maximize its adversarial effect. The pioneering work by Brown et al. [19], "Adversarial Patch" is a quintessential example of a white-box attack. In contrast, a black-box attack scenario presents a more realistic and challenging condition where the attacker has no internal knowledge of the model and can only interact with it by querying its API and observing its outputs, such as confidence scores or final labels. Black-box attacks can be further subdivided into query-based and transfer-based approaches. Query-based methods iteratively refine the patch by sending numerous queries to the model to estimate the gradient or search the space of possible perturbations [43–45]. Transfer-based attacks, on the other hand, leverage the transferability of adversarial examples; an attacker first creates a patch against a locally trained surrogate model, with the expectation that the patch's adversarial properties will transfer to the unknown target model [46,47].

2.3.2. Attack Goal

The second dimension concerns the attacker's intended outcome on the model's output. Attacks are broadly categorized as either targeted or non-targeted [19]. A targeted attack aims to precisely control the model's misclassification, forcing it to predict a specific class chosen by the adversary. For instance, the objective may be to cause a model to recognize a stop sign as a speed limit sign. This requirement for specificity makes targeted attacks generally more difficult to execute successfully. Conversely, a non-targeted attack seeks only to cause any form of misclassification, without specifying the erroneous class. The goal is simply to degrade the model's accuracy, such as causing a model to classify a cat as any class other than cat, which is a less constrained and often more easily achievable objective [48].

2.3.3. Patch Properties

The third dimension focuses on the physical and spatial characteristics of the patch itself. A fundamental distinction is drawn between digital-world patches and physical-world patches. Digital patches are applied directly to the pixel space of an image, which simplifies experimentation and is common in foundational research. Physical-world patches, however, are rendered into real-world objects, such as stickers or posters, and are captured by a camera under varying conditions. This setting introduces significant challenges, including changes in lighting, viewpoint, distance, and camera resolution, necessitating robust patch generation that can withstand these transformations [49]. Beyond the digital-physical dichotomy, patch properties also encompass shape and location. While early research predominantly utilized simple shapes like squares and circles, recent studies have explored arbitrary shapes to enhance stealth or efficacy. Furthermore, the patch location can be either fixed within the image frame or treated as an optimizable parameter during the attack generation process to identify the most vulnerable region for a given target [50–52].

2.3.4. Application Scenario

The final dimension categorizes attacks based on the computer vision task they are designed to subvert. The complexity and nature of the attack vary significantly across different tasks. Image classification, the most studied scenario, involves causing the model to misclassify the entire image [29]. Attacking object detection models is more complex, with objectives that can be categorized into three subtypes: vanishing attacks, which aim to hide an object from the detector; fabrication attacks, which cause the detector to perceive a non-existent object; and misclassification attacks, which change the label of a correctly detected object. Other critical application scenarios include face recognition, where attacks aim to deceive verification or identification systems, and a growing body of work targeting other tasks such as image segmentation and video action recognition [53,54]. Each scenario presents unique constraints and challenges that shape the design of effective patch-based attacks.

 doi:10.20944/preprints202510.1706.v1

2.4. Comparison with Other Attack Types

Adversarial patch attacks differ in important ways from other commonly studied adversarial threats in machine learning, including imperceptible perturbation attacks, data poisoning attacks, and backdoor attacks.

Imperceptible perturbation attacks generate small, pixel-level modifications that are distributed across the entire input image. These perturbations are usually constrained by an L_p-norm bound to ensure that they remain visually indistinguishable from the original image, while still being able to significantly alter the model's prediction [24]. In contrast, patch attacks rely on localized and often visible modifications, which can be optimized to achieve universal or physically realizable effects without requiring pixel-level stealth.

Data poisoning attacks compromise the training process by injecting carefully crafted malicious samples into the training dataset. A model trained on such contaminated data inherits hidden vulnerabilities, such as reduced generalization or susceptibility to specific triggers [18]. Unlike poisoning attacks, patch attacks operate purely at inference time and do not require access to the training pipeline. Their practicality stems from the fact that they can be deployed directly on test-time inputs.

Backdoor attacks, on the other hand, embed hidden triggers into a model during training such that the presence of the trigger in an input forces the model to output an attacker-specified label [55]. While backdoor triggers can be considered as a special form of adversarial perturbation, they differ from patch attacks in that they require control over training or fine-tuning to implant the trigger. Patch attacks, in contrast, are training-agnostic and exploit the vulnerabilities of pre-trained models by applying adversarial patterns externally.

Thus, although imperceptible perturbation, poisoning, and backdoor attacks each present significant risks, patch attacks are uniquely concerning due to their visibility, ease of deployment, and robustness in the physical world. These characteristics make patch attacks a distinct and urgent area of study within adversarial machine learning.

3. Taxonomy of Patch Attack Methods

Patch attacks can be systematically categorized according to the downstream vision task they target. Unlike imperceptible perturbations that are typically designed to affect any input image in a subtle manner, patch-based perturbations often exploit the spatially localized vulnerabilities of deep neural networks. As such, their design and impact differ depending on whether the targeted model is built for image classification, object detection, or other vision-related tasks. In the following, we discuss representative patch attack strategies within these three categories.

3.1. Patch Attacks in Image Classification

The earliest and most widely studied application of adversarial patches is in image classification. In this setting, a carefully crafted patch is embedded into an image such that the model consistently misclassifies the input regardless of the original content. A typical example is the universal adversarial patch proposed by Brown et al. [19], which can force a wide range of classifiers to predict a targeted label with high confidence. Subsequent works extended this idea to improve transferability across models, enhance robustness under physical-world conditions (e.g., printing and photographing patches), and adapt patches to specific regions of interest such as faces or clothing. Karmon et al. [50] propose adversarial patches that are localized, visible, and restricted to only 2% of image pixels, avoiding the main salient object. The learned patches often resemble features of the target class, such as textures, body parts, or global shapes, despite their small size. Gradient analyses show that the patches are not salient to the network, indicating the model is fooled without "noticing" the perturbation. This work highlights a significant security risk and opens avenues for studying architectural vulnerabilities and the inner workings of deep classifiers.

Liu et al. [26] propose perceptual-sensitive GAN (PS-GAN) to generate adversarial patches with both strong attacking ability and high visual fidelity. Unlike prior approaches, PS-GAN leverages

 doi:10.20944/preprints202510.1706.v1

the perceptual sensitivity of the attacked network, ensuring patches remain visually natural and context-correlated. The framework adopts a patch-to-patch translation process, allowing adversaries to create diverse styles of adversarial patches from arbitrary seed patches. An attention mechanism is integrated to capture spatial sensitivity and guide patch placement, thereby enhancing attack effectiveness. Following this direction of improving visual fidelity and placement, Zhou et al. [56] propose DiAP, which is a data-independent adversarial patch method that generates transferable patches without any access to the model's training data. It first performs non-targeted attacks by disrupting features at multiple network layers, then leverages those perturbations to construct targeted patches. Meanwhile, Kang et al. [27] introduce adversarial image patches (AIPs) designed with optimized location, size, and perturbation ratio to mislead both deep neural networks (DNNs) and their interpretation models. The authors propose a general framework for adversarial Grad-CAM, demonstrating scenarios where patches deceive classification models while also altering heatmap explanations. Their method generates highly localized patches, covering only 1.5–3

Extending the patch idea to specialized domains, Wang et al. [57] propose a Multi-Patch Adversarial Attack (MPAA) method to address the challenge of adversarial patch attacks in remote sensing image (RSI) scene classification. Instead of relying on a single large patch, MPAA deploys multiple small patches at key locations through a constrained optimization framework combining location selection and patch optimization. Building on multi-patch strategies for RSIs, Huang et al. [58] propose DeMPAA, a deployable multi-mini-patch adversarial attack tailored for remote sensing image (RSI) classification that places multiple small patches at key feasible locations rather than a single large patch. DeMPAA formulates joint optimization of patch locations and patch patterns as a constrained optimization problem and solves it with a two-stage framework: a Feasible and Effective Map Generation (FEMG) module to exclude infeasible locations and score location effectiveness, and a Patch Generation (PG) module that uses probability-guided random sampling (PRSamp) to select locations and gradient-based optimization of patches. Taking a different geometric approach, Ran et al. [39] introduce a Cross-Shaped Patch Attack (CSPA), which departs from conventional rectangular or grid-like patches by adopting two thin, long, and perpendicular line segments intersecting at the midpoint of the image. This design enables the patch to extend toward the four image corners, thereby achieving a more global perturbation effect while maintaining continuity. To optimize both the content and placement of the patch, the authors employ a random search-based algorithm tailored for black-box attack scenarios. Finally, moving toward practical and stealthy deployment, Tiliwalidi et al. [59] propose a novel, black-box camera-patch physical attack that uses a single, easily deployed patch placed on the camera lens to produce stealthy adversarial inputs for DNN-based perception. The patch is optimized with particle swarm optimization (PSO) to maximize attack effectiveness while maintaining minimal visual conspicuity and deployment complexity. Overall, classification-task patch attacks are characterized by their generality, high success rate, and relative simplicity in design.

3.2. Patch Attacks in Image Detection

In contrast to classification, object detection models present unique challenges for adversarial patch attacks because they simultaneously localize and classify multiple objects. According to this, patch attacks often aim to either conceal the presence of a target object (evasion) or generate false detections (fabrication). For example, stop sign attacks against autonomous driving systems demonstrate that strategically placed stickers can cause detectors like YOLO or Faster R-CNN to fail to recognize critical road signs. More advanced approaches adaptively optimize patch positions or leverage physical transformations to ensure effectiveness under diverse viewing angles and lighting conditions. For instance, Lee et al. [60] propose a physical adversarial patch that, when placed anywhere in a scene, universally suppresses object detections rather than needing to overlap target objects. The attack is designed to disrupt YOLOv3's feature extraction so that virtually all objects in the image are missed, including those far from the patch. They validate the method quantitatively on COCO (measuring mAP degradation) and qualitatively in a real-time physical scenario using a webcam. This

 doi:10.20944/preprints202510.1706.v1

work therefore introduces a new threat model for object detectors: a single, portable sign-like patch can blind a detector to all objects in a scene without modifying those objects.

Wu et al. [61] propose Diffused Patch Attack (DPAttack), which introduces asteroid-shaped or grid-shaped diffused patches that alter only a small number of pixels. Unlike traditional patch or pixel-level perturbations, these diffused patches influence a wider range of features in the detector's feature maps while maintaining high efficiency. They design a tailored attack loss that emphasizes unsuccessfully attacked proposals and reduces false positives. Building on the theme of targeted and efficient perturbations for detectors, Huang et al. [62] introduce a novel patch-based attack method, RPAttack to target general object detectors such as YOLO v4 and Faster R-CNN. The approach employs a patch selection and refining scheme to dynamically identify key pixels and gradually eliminate inconsequential perturbations. To ensure stable performance across models, the method balances gradients between detectors during training, preventing over-optimization of a single model. In a related direction that emphasizes adaptability to changing viewpoints, Hoory et al. [63] introduce a Dynamic Adversarial Patch method for attacking object detectors in real-world settings. Unlike static patches, this approach employs multiple pregenerated adversarial patches placed on the target object and dynamically switches between them based on the camera's position. The implementation uses flat screens to display and change patches in real time, enabling adaptability to varying viewpoints and nonplanar surfaces such as cars.

Continuing the exploration of detector-targeting patches, Wang et al. [64] propose a novel adversarial patch attack that hides objects of a target class from state-of-the-art object detectors by embedding a crafted patch on the object. The patch is generated by minimizing a specially designed detection score; three interpretable variants of this score are introduced to quantify and suppress detector outputs for specific objects. The optimization yields highly transferable patches that successfully fool multiple detection architectures and datasets in digital experiments, achieving a minimum recall of 11.02% and a maximum fooling rate of 81.00%. Physical-world feasibility is demonstrated by transferring the learned patch to a portable display and evading a real-time surveillance detector. Complementing optimization-based attacks with feature-guided approaches, Lang et al. [65] propose an attention-guided adversarial-patch algorithm that extracts key vehicle feature vectors to compute a feature-aggregation field and steer both the generation and placement of patches. The method optimizes the patch (and its attachment mechanism for position and size) to account for real-world variations—illumination, distance, angle, color and resolution—so that camera and mobile-captured images will reduce detector confidence and background/target discrimination. By minimizing a tailored loss function the generated patches can hide or cause misclassification of vehicles, and patches produced for one detector transfer effectively to others.

Shifting focus toward remote sensing and aerial domains, Sun et al. [66] introduce a novel and more threatening patch attack (TPA) for object detection in optical remote sensing images (O-RSIs) without sacrificing visual quality. To address inconsistencies in existing patch selection, the authors propose a first-order difference (FOD) based scheme that selects subpatches by comparing the objective function before and after masking. In addition, they design a bounding box drifting loss (BDL), an IoU-based objective function that mitigates gradient inundation by pushing detected boxes away from the originals until no overlap remains. Extending the idea of layer- or intermediate-output–based optimization, Tang et al. [67] propose a novel adversarial patch attack method for aerial imagery object detectors by designing a loss function based on intermediate outputs rather than the final detection head. Extensive experiments on DOTA, RSOD, and NWPU VHR-10 datasets demonstrate that this approach significantly improves attack effectiveness compared to prior methods. The study further shows that objectness scores are more effective than class scores for optimizing patches and analyzes how parameters such as patch size, position, and number influence attack performance. In addition, ensemble training is introduced to enhance the transferability of adversarial patches across different datasets and models.

 doi:10.20944/preprints202510.1706.v1

Parallel to these domain-specific strategies, Deng et al. [68] propose a rust-style adversarial patch generation framework that uses style transfer to create natural, camouflaged patches designed to evade object detectors in remote sensing images. The method leverages heat-map based interpretability to identify key recognition regions and generates irregular-shaped, small-area patches to minimize visual suspicion while maximizing impact on detectors. To improve real-world effectiveness, they apply physical-domain augmentations (rotation, scaling, brightness, etc.) during training to increase robustness against imaging variations and evaluate attacks against YOLOv3. Addressing 3D and autonomous-driving contexts, Wang et al. [69] introduce a unified framework for generating physically printable adversarial patches tailored for 3D object detection in autonomous driving. The framework supports two attack goals: instance-level hiding, where patches pasted on vehicles cause them to evade detection, and scene-level creating, where patches placed in the scene induce false object detections. To enable effective patch learning in 3D space, the authors design a differentiable image-3D rendering algorithm, combined with a Sparse Object Sampling Strategy to preserve perspective realism and a Patch-Oriented Adversarial Optimization to focus training on patch regions.

Focusing on UAV-specific challenges, Shrestha et al. [53] propose a novel, robust adversarial-patch generation scheme specifically designed for UAV settings by modeling camera perspective, viewing angle, distance and brightness variations during patch creation. The resulting patches are effective at degrading object detectors even when models differ in initialization or architecture, demonstrating strong cross-model robustness. Exploring triggerable and multi-modal mechanisms, Zhu et al. [70] propose TPatch, a physical adversarial patch that remains benign under normal conditions but can be triggered by acoustic signal-injection attacks on cameras to launch hiding, creating, or altering attacks. To design TPatch the authors introduce a trigger-oriented optimization for attack effectiveness, a content-based camouflage scheme to reduce human suspicion, and an attack-robustness enhancement to improve real-world practicality. They evaluate TPatch in simulation and outdoor driving tests against three object detectors (YOLOv3/v5, Faster R-CNN) and eight image classifiers in both white-box and black-box settings. Wei et al. [71] propose a camera-agnostic physical adversarial patch (CAP) attack that explicitly models the camera's role in the physical-to-digital pipeline. Central to their method is a differentiable camera Image Signal Processing (ISP) proxy network that closes the gap between real-world capture and digital images and is integrated into the attack pipeline. They formulate a zero-sum adversarial game where the attack module optimizes patches to maximize detector failure while the ISP proxy's conditional parameters are adversarially optimized to minimize attack effectiveness, improving cross-camera stability. Real-world experiments on two cameras (Sony, Canon) and four smartphones (iPhone, Redmi, Huawei, Samsung) show the CAP attack yields stronger, more reproducible person-concealment across diverse imaging hardware.

In a different modality-focused vein, Zhang et al. [72] introduce CAPatch, a physical adversarial patch that manipulates image captioning outputs by generating irrelevant sentences or suppressing key terms. The method incorporates detection assurance and attention enhancement to amplify adversarial impact, along with robustness improvements to withstand distortions from physical printing and recapturing. By exploiting both the feature extraction and description generation stages, CAPatch demonstrates the feasibility of disrupting multi-modal captioning pipelines. Experimental results across several captioning models in both digital and physical settings highlight the practicality and generalizability of this approach under diverse environmental conditions. Exploring decision-based and video scenarios, Jiang et al. [73] introduce a novel attack scenario—decision-based patch attacks on video models—which combines patch attacks and decision-based settings to assess robustness in video recognition. To address the challenges of high parameter space and limited feedback, the authors propose Spatial-Temporal Differential Evolution (STDE). STDE adaptively selects keyframes via temporal difference, embeds target videos as patch textures, and optimizes with spatial-temporal mutation and crossover to minimize patch area. Experiments on UCF-101 and Kinetics-400 show that STDE achieves state-of-the-art fooling rates with fewer queries, smaller patches, and strong imperceptibility.

 doi:10.20944/preprints202510.1706.v1

Shifting to non-visible-spectrum and inspection domains, Liu et al. [74] proposes X-Adv, a method that generates physically printable metal objects with adversarial shapes (rather than textures) designed to fool X-ray prohibited-item detectors under texture-fading conditions. To enable shape optimization for the X-ray domain, the authors introduce a differentiable converter that maps gradient signals from a surrogate detector to 3D-printable adversarial geometries, avoiding reliance on visible-spectrum textures. To handle complex clutter and heavy occlusion in luggage, they use a policy-based reinforcement learning strategy to discover robust, worst-case placement locations for the printed adversarial objects. The pipeline is validated with extensive digital experiments and real-world tests on a commercial X-ray inspection system, and the authors release the XAD physical-world X-ray adversarial attack dataset.

Complementing GAN- and shape-based strategies, Agrawal et al. [75] propose a black-box transferable patch attack (TPA) that uses a GAN, with a generator and discriminator, to create imperceptible adversarial patches added to face images. Patches are optimized to be visually invisible while maximizing transferability across models. Wang et al. [54] propose Sensitive Region Patches, a simple physical attack that alters an object's infrared signature by attaching low-cost aerogel insulation patches to learned sensitive regions on pedestrians and cars to mislead detectors. A novel consensus selection strategy automatically identifies those sensitive regions by aggregating importance across multiple detection models, avoiding manual region design. They validate the method in both digital and physical experiments across angles, distances, postures, and scenes, achieving over 70% attack success rate (ASR).

Similarly targeting infrared modalities, Wei et al. [76] propose infrared adversarial patches, physically realizable patches made of thermal-insulating material that are attached to targets to manipulate their thermal signature and fool infrared object detectors. To optimize the patch shape and placement jointly, they introduce a novel aggregation regularization that guides simultaneous learning of location and geometry, enabling a simple gradient-based optimization. The method is highly practical: patches can be manufactured in about 0.5 hours and achieve $>$90% attack success rate against pedestrian and vehicle detectors in physical tests across different angles, distances, poses, and scenes. Chen et al. [77] propose the Latent Diffusion Patch (LDP), a novel adversarial-patch method that uses a pretrained encoder to compress natural images into a perceptual feature space and then trains a diffusion model on those features. LDP generates adversarial patterns by exploring the diffusion model's latent space and applying image-denoising techniques to iteratively polish patches so they blend with natural image statistics. By constraining the variation range of latent variables during generation, LDP forces patch feature vectors to remain close to the latent distribution of real images, producing visually plausible, camouflaged patches.

Complementing latent-space approaches with uncertainty-driven objectives, Lin et al. [78] propose an entropy-boosted loss that directly increases class-probability uncertainty from an object detector to drive adversarial patch optimization. The loss integrates entropy with the detector's predicted class probabilities so the learned patch shifts detections away from the "person" class. Applied to YOLOv2, YOLOv3 and YOLOv4, this objective consistently produces patches that the detectors interpret as a benign object (e.g., "potted plant"), thereby concealing pedestrians. Liu et al. [21] propose EAP, an effective black-box impersonation attack that generates adversarial patches for fooling face-recognition systems in the physical world. EAP creates printable patches (suitable for mobile/compact printers) that can be attached to a source face to induce targeted impersonation. To boost transferability, the method applies random similarity transformations and an image-pyramid strategy to increase input diversity, and uses a meta-ensemble attack that aggregates gradient features from multiple pre-trained face models. Zhou et al. [79] propose a Dual-Perception-Based Framework (DPBF) that generates the More Vivid Patch (MVPatch) to jointly improve adversarial patch transferability, stealthiness, and practicality. The Model-Perception-Based Module (MPBM) uses an ensemble strategy grounded in generalization theory to reduce object confidence across multiple detectors, boosting transferability and stabilization compared with single-model attacks. The Human-Perception-Based Module (HPBM)

 doi:10.20944/preprints202510.1706.v1

enforces visual similarity with real images via a lightweight perceptual regularizer, producing natural, inconspicuous patches without relying on additional generative models.

Advancing GAN-based latent-space search methods, Wang et al. [80] introduce a GAN-based framework for generating adversarial patches against object detection models. The method leverages dataset slicing to train a GAN that learns background visual features, from which adversarial patches are searched by exploring the latent space. These patches achieve strong attack efficacy and maintain high environmental consistency. Addressing pose and style consistency jointly, Zhou et al. [81] introduce a novel approach to synthesizing adversarial patches that appear visually natural in both pose and texture. A PosePatch network is proposed to adapt patches to human poses through perspective transformation, while a StylePatch network harmonizes patch textures with image content. These two components are jointly trained in an end-to-end manner to generate effective and inconspicuous adversarial patches. Finally, exploring triggered physical patches and traffic-sign scenarios, Yuan et al. [82] propose ITPATCH, an invisible, triggered physical adversarial patch that uses carefully designed fluorescent-ink perturbations to create robust adversarial examples for traffic-sign recognition (TSR) systems. The patch remains stealthy under normal lighting and is activated by invisible ultraviolet light, at which point the fluorescent pattern causes misclassification or detection failure.

In a similar spirit of creative physical realizations, Hu et al. [83] propose LMBC (Leaf-like Mask Bar Code), which is a novel physical black-box adversarial patch designed for multi-angle evasion of infrared vehicle detectors, using a leaf-inspired mask to constrain the patch contour and improve environmental adaptability. The method physically implements leaf-like structures with readily available infrared-coating materials (aluminum sheet and kraft paper) to produce high-resolution, camera-visible adversarial patches. Adversarial parameters, rotation angle, sparsity, and position, are jointly optimized with a Genetic Algorithm with Multi-segment (GAM) to maximize attack robustness across viewing angles. Liu et al. [84] proposed RPAU, a robust physical attack framework against UAVs that directly compromises flight safety through three attack modes: Hiding Attack (HA), Yaw Attack (YA), and Obstacle Attack (OA). The framework addresses major design challenges by introducing a nested patch for continuous perturbation, extended image transformations to reduce digital–physical discrepancies, and a time-dependent mechanism for perturbation optimization. Experiments were carried out in digital, simulation, and real-world physical domains. The results demonstrate that RPAU achieves a substantially higher attack success rate than baseline methods, remaining effective even in the physical world. Attacks on detection tasks highlight the severe real-world risks of patch-based adversarial examples, especially in safety-critical systems.

3.3. Patch Attacks on Other Vision Tasks

Beyond classification and detection, patch attacks have been explored in a variety of other vision-based tasks, reflecting the versatility of this threat model. In face recognition and authentication, adversarial patches in the form of eyeglass frames or facial accessories have been shown to enable impersonation and dodging attacks [48]. In semantic segmentation, patches can cause widespread mislabeling of regions, undermining applications such as medical image analysis and autonomous navigation [85]. Additionally, research has extended patch attacks to pose estimation, visual tracking, and even cross-modal systems that integrate vision with natural language. More vision task patch attack methods is provided in Table 1. These studies demonstrate that patch-based adversarial methods pose a broad and evolving challenge, threatening the reliability of numerous AI-powered vision systems.

 doi:10.20944/preprints202510.1706.v1

Table 1. Overview of Patch Attacks on Other Vision Tasks.

Task Category	Ref. & Year	Method / Approach	Key Characteristics	Performance Metrics
Image Segmentation	Ref. [85], 2022	Extends the Expectation Over Transformation (EOT) paradigm to semantic segmentation. Adversarial patches are printed on billboards and deployed in outdoor driving experiments.	Semantic segmentation (SS), digital and physical attack.	mIoU (mean Intersection-over-Union), ASR(Attack Success Rate)
Face recognition	Ref. [48], 2016	White-box optimization of adversarial patterns. Differentiable end-to-end pipeline through the face-recognition model.	Physically realizable (printed, wearable eyeglass frames). Inconspicuous.	ASR, Dodging results, Impersonation results
Remote sensing	Ref. [86], 2024	Self-supervised harmonization module integrated into patch generation. Aligns patch appearance with background imaging environment.	digital-to-physical visual inconsistency. Self-supervised. Harmonization-guided optimization	ASR, FLOPs
Object Tracking	Ref. [87], 2019	Optimize visually inconspicuous poster textures. Apply Expectation Over Transformation (EOT) for physical robustness.	Inconspicuous / natural-looking textures. Works against regression-based trackers	Success rate of evasion, Visual imperceptibility
Optical Flow Estimation	Ref. [88], 2019	Extend adversarial patch attacks to optical flow networks. Analyze encoder–decoder vs. spatial pyramid architectures.	Patch attacks propagate errors beyond attack region. Patches can distort object motion	End Point Error (EPE), relative degradation.
Crowd Counting	Ref. [89], 2022	Perceptual Adversarial Patch (PAP) framework. Uses adaptive crowd density weighting to capture invariant scale features.	Model-shared perceptual features. Effective in both digital and physical world.	Mean Absolute Error (MAE). Mean Squared Error (MSE).
X-ray Detection	Ref. [74], 2023	Shape-based (not texture) adversarial generation to handle X-ray color/texture fading. Policy-based reinforcement learning to find worst-case placements inside luggage under heavy occlusion.	Texture-free, geometry-driven adversarial agents (metal objects). Designed for physical realizability (3D-printable).	ASR, Detection accuracy / mAP drop

 doi:10.20944/preprints202510.1706.v1

4. Defense Methods Against Patch Attacks

Given the significant risks posed by adversarial patches, a wide range of defense strategies have been proposed to mitigate their impact on vision systems. Unlike imperceptible perturbations, patch attacks are visually localized and often conspicuous, which opens opportunities for defenses that exploit their spatial and structural characteristics. Existing approaches can be broadly grouped into three major categories: Patch Localization and Removal-based Defenses; Input Transformation and Reconstruction-based Defenses; Model Modification and Training-based Defenses. In this section, regarding defense methods, we focus particularly on works from 2023 to 2025. In the following, we discuss each defense category in detail, highlighting representative works, their core methodologies, and their strengths and limitations.

4.1. Patch Localization and Removal-Based Defenses

One intuitive and widely adopted strategy against adversarial patch attacks is to directly localize and remove the malicious patch region before feeding the input into the model. Since adversarial patches often exhibit visually or statistically abnormal patterns compared to natural regions, localization methods rely on cues such as entropy, saliency, or explainability. Once detected, the suspicious region is either masked, replaced, or suppressed to mitigate its influence. Representative approaches include entropy-based localization, such as Jedi [90] (Tarchoun et al., 2023), which addresses adversarial physical patches from an information-theoretic perspective. It first employs entropy analysis to identify potential patch regions, leveraging the observation that adversarial patches exhibit high entropy even when naturalistic. To further localize patches, Jedi integrates an autoencoder that reconstructs high-entropy regions, enabling accurate patch completion. Similarly, PatchZero [91] (Xu et al., 2023) proposes a general defense pipeline against white-box adversarial patches without requiring retraining of the downstream model. The method detects adversarial regions at the pixel level and neutralizes them by repainting with mean pixel values. To enhance robustness, a two-stage adversarial training scheme is further incorporated to resist adaptive attacks. ObjectSeeker [92] (Xiang et al., 2023) introduces a defense framework for object detectors against patch hiding attacks. Its core idea is patch-agnostic masking, which removes adversarial effects without requiring prior knowledge of the patch's shape, size, or location. This enables standard object detectors to operate reliably on masked images. Moreover, ObjectSeeker incorporates a certification procedure that provides formal guarantees of robustness under white-box adaptive attacks.

Similarly, PAD [93] (Jing et al., 2024) introduces a patch-agnostic defense method designed to localize and remove adversarial patches without relying on prior knowledge or additional training. It leverages two inherent characteristics of adversarial patches, semantic independence and spatial heterogeneity, to detect patch regions effectively. Unlike existing approaches that depend on attack data, PAD maintains compatibility with any pre-trained object detector. Bunzel et al. [94] propose a novel detection method based on edge detection. The key idea is that adversarial patches typically form high-entropy regions with dense edges and fine details, which can be effectively identified. Hofman et al. [95] introduce X-Detect, which is a novel adversarial patch detection framework designed for object detection models. It employs an ensemble of explainable-by-design detectors that leverage object extraction, scene manipulation, and feature transformation to identify adversarial samples. The method can detect attacks in real time, provide interpretable explanations for alerts, and generalize to unseen threats. Wu et al. [96] propose NAPGuard, a detection framework designed to counter naturalistic adversarial patches (NAPs). To improve precision, it employs aggressive feature aligned learning with a pattern alignment loss, enabling the model to capture more accurate aggressive patterns despite deceptive appearances. To enhance generalization, it introduces natural feature suppressed inference, which mitigates disturbances from diverse NAP representations through a unified feature shield module.

More recent frameworks, such as Saliuitl [97] (Victorica et al.,2025), a recovery method against adversarial patches that is independent of the number, shape, and contiguity of patches. Saliuitl

 doi:10.20944/preprints202510.1706.v1

first detects the presence of patch attacks using an ensemble of binarized feature maps generated by multiple saliency thresholds. Once detected, the method recovers clean predictions by localizing and inpainting the adversarial patches guided by the feature map ensemble. Overall, this class of defenses benefits from interpretability and simplicity, as they aim to neutralize adversarial effects at the input level. However, they can be challenged by stealthy or highly camouflaged patches that avoid detection.

4.2. Input Transformation and Reconstruction-based Defenses

Another line of research aims to reconstruct a purified version of the input to suppress the effect of adversarial patches. Instead of explicitly detecting the malicious region, these approaches exploit generative or diffusion-based models to re-synthesize an image that retains benign content while removing adversarial noise. For example, Diffender [99] (Kang et al., 2024) introduce a diffusion-based defense framework designed to counter adversarial patch attacks. It leverages the Adversarial Anomaly Perception (AAP) phenomenon, which enables the diffusion model to detect and localize adversarial patches by exploiting distributional discrepancies between natural and adversarial regions. A single text-guided diffusion model is employed to jointly perform patch localization and restoration, where accurate localization improves restoration and successful restoration, in turn, refines localization. To enhance adaptability, DIFFender incorporates vision-language pre-training and a few-shot prompt-tuning algorithm, allowing efficient tuning without extensive retraining. The framework also integrates three specialized loss functions to co-optimize localization and restoration, improving robustness against diverse patch attacks. Extensive evaluations on classification, face recognition, and physical-world settings show that DIFFender achieves strong generalization across scenarios, classifiers, and attack methods.

Similarly, Wei et al. [102] (2025) identify a novel phenomenon called Adversarial Anomaly Perception (AAP), which enables adversarial patch localization by analyzing discrepancies among multiple denoised images. Building on this insight, the authors propose DIFFender, a diffusion-based defense framework that integrates patch localization and restoration within a unified model. The localization process guides targeted restoration, while restoration feedback refines localization, creating a synergistic defense pipeline. To enhance robustness, DIFFender incorporates text-guided diffusion models with a few-shot prompt-tuning strategy for efficient adaptation to defense tasks. Furthermore, DIFFender is extended to the infrared domain, addressing domain shift and weaker texture challenges through an Infrared Domain Constrained (IDC) token and specialized loss functions. These defenses have the advantage of being attack-agnostic, as they do not rely on prior knowledge of patch location or shape. However, they introduce additional computational overhead due to the iterative nature of generative reconstruction, which may hinder deployment in real-time systems.

 doi:10.20944/preprints202510.1706.v1

Table 2. Representative Patch Defense Methods in Computer Vision.

Ref. & Year	Defense Category	Method / Approach	Key Characteristics
Ref. [93], 2024	Detection-based	Remove adversarial patches by leveraging their semantic independence and spatial heterogeneity	Patch-agnostic, Training-free, Effective across modalities
Ref. [95], 2024	Detection-based	Uses an ensemble of explainable detectors to spot inconsistencies from adversarial patches and raise alerts.	Real-time detection, Explainability, Generalization to new attacks
Ref. [98], 2023	Certified Defenses	Divide the binary into byte chunks and make the final decision by majority voting over chunk predictions.	Chunk-Based Smoothing, Certifiable Robustness, Majority Voting
Ref. [96], 2024	Detection-based	It improves robustness to naturalistic adversarial patches by leveraging their aggressiveness and naturalness.	Targets naturalistic adversarial patches, Feature-level modulation, Improves precision and generalization
Ref. [97], 2025	Detection-based	Detect adversarial patches using a binarized feature map ensemble generated with multiple saliency thresholds	Explicit patch detection, inpainting-based recovery, Low computational complexity
Ref. [99], 2024	Detection-based & Pre-processing	A text-guided diffusion model detects and localizes adversarial patches by identifying distributional differences, then restores the image to remove the perturbations.	Utilizes diffusion models, few-shot tuning
Ref. [100], 2024	Detection-based & Pre-processing	Analyze the visual and feature-level inconsistencies introduced by adversarial patches to locate and filter out adversarial regions	Dual Attack Resistance, High Generalization
Ref. [101], 2025	Detection-based	Apply randomized Fourier-space sampling masks to enhance robustness to occlusion and adversarial perturbations. SAF (Split-and-Fill) Strategy	Fourier-based augmentation, edge-aware segmentation, and adaptive reconstruction.

 doi:10.20944/preprints202510.1706.v1

4.3. Model Modification and Training-based Defenses

Instead of manipulating the input, another category of defenses focuses on increasing the robustness of the model itself through architectural modifications, robust training, or adaptive mechanisms. The underlying principle is to make the neural network inherently resistant to adversarial patches, even when they are present in the input. For instance, HARP [103] (Cai et al., 2023) proposes a Hyperplasia-based Adversarial Patch Defense (HARP), inspired by the biological phenomenon of bone hyperplasia. HARP introduces lightweight hyperplasia modules with residual structures and attention mechanisms, which are inserted into key areas of the original detector without altering its existing weights. These modules are trained via adversarial training to enhance robustness against adversarial patches while maintaining clean performance. HARP is a universal model-side defense applicable to various object detectors such as YOLOv3 and SSD, with low training cost due to the limited number of trainable parameters.

Jujutsu [104] (Chen et al., 2023) introduces a defense framework designed for both detecting and mitigating adversarial patch attacks. For detection, it leverages the insight that adversarial patches generate localized, input-agnostic features with dominant influence on model predictions. It identifies the potential patch region using saliency maps and applies a robust preprocessing step to highlight adversarial rather than benign influential features. To distinguish adversarial patches from benign ones, Jujutsu employs guided feature transplantation, transferring the extracted patch to a low-saliency region in a new input and testing its effect on classification. For mitigation, Jujutsu reconstructs the corrupted patch region using GANs, exploiting the unperturbed pixels to recover clean semantic content for robust predictions. Finally, it introduces a parametric strategy that enables configurable trade-offs between detection accuracy and false positive rate depending on application needs. Yu et al. [105] propose a novel defense against universal adversarial patch attacks by analyzing their impact on deep feature representations. They reveal that adversarial patches cause abnormally large feature norms concentrated at the patch location, which can dominate pooled features in classifiers or suppress objectness scores in detectors. To mitigate this effect, they introduce the Feature Norm Suppressing (FNS) layer, which restricts feature norms above a threshold using non-increasing functions such as clipping, exponential decay, or Gaussian decay. The method can be flexibly inserted into CNN architectures, including ResNet and GoogLeNet, without significant computational overhead. Lin et al. [100] propose NutNet, a lightweight reconstruction-based autoencoder designed to detect adversarial patches without relying on pre-generated patches. NutNet is trained solely on clean samples, treating them as in-distribution data, while adversarial patches are considered out-of-distribution. Detection is achieved by measuring the reconstruction error, as clean images can be faithfully reconstructed but adversarial patches cannot. To enhance robustness, NutNet introduces Image-splitting, which divides inputs into blocks to magnify the distinction between patched and clean regions. It further employs Destructive Training, deliberately restricting the decoder's generative capacity so that only normal images can be reconstructed.

More recent frameworks, such as Radap [101] (Liu et al., 2025), a robust and adaptive defense framework designed to counter adversarial patch attacks in both closed-set and open-set face recognition systems. It employs a patch segmenter that detects adversarial patches and conceals them using masking techniques. To improve detection, F-patch generates diverse adversarial patches via Fourier-space sampling, enabling the segmenter to recognize patches of various shapes. An edge-aware binary cross-entropy (EBCE) loss is introduced to enhance boundary detection accuracy. To strengthen occlusion robustness, FCutout applies random Fourier-space masks as a data augmentation strategy. Finally, the split-and-fill (SAF) strategy mitigates the vulnerability of the segmenter to adaptive white-box attacks, and experiments confirm RADAP's superior defense performance over state-of-the-art methods. Overall, compared to input-level defenses, model modification approaches are generally more robust and integrated, but they often require retraining or architectural changes, which may limit their applicability to pretrained or large-scale models.

 doi:10.20944/preprints202510.1706.v1

4.4. Task-Specific / Domain-Specific Defenses

In addition to general defense approaches, researchers have introduced task-specific and domain-oriented strategies aimed at the unique vulnerabilities present in different applications. Because adversarial patches often exploit task-dependent visual cues, designing defenses that are tailored to the characteristics of the target system can lead to greater robustness and practical effectiveness. For example, Zheng et al. [106] analyze why adversarial attacks are effective against RGB-D systems and introduces a detection-based defense. The defense compares each input's RGB-D representation with the centroid of the predicted class, flagging samples as adversarial if the distance exceeds a threshold. Unlike adversarial training, this method does not rely on incorporating adversarial examples during training, yet it improves robustness against both standard and adaptive attacks. Chattopadhyay et al. [107] propose Outlier Detection and Dimension Reduction (ODDR), a model-agnostic defense that detects adversarial patches by identifying clusters of outliers in input features. By applying localized dimension reduction around the detected outlier region, ODDR neutralizes patch effects while retaining essential information.

Regarding to the autonomous driving, Liang et al. [108] introduce a novel adversarial patch attack for AV visual object detection, designed with both evasion and misclassification modes through dedicated optimization. To counter this, they propose a defense method that leverages texture features to detect adversarial patch regions and applies local denoising for mitigation. Experiments on the KITTI dataset and real driving scenes show that the attack can significantly degrade detection accuracy for cars and pedestrians. Similarly, Chattopadhyay et al. [109] introduce a model-agnostic defense mechanism against adversarial patch attacks by treating patches as anomalies in the image distribution. The approach employs a clustering-based technique (DBSCAN) to isolate anomalous image segments through a three-stage pipeline of Segmenting, Isolating, and Blocking. Strack et al. [110] investigate defense strategies against adversarial patch attacks on infrared human detection and introduces Patch-based Occlusion-aware Detection (POD). The method augments training samples with random patches and enables the model to both detect people and localize adversarial patches. Wu et al. [111] propose a real-time method for generating imperceptible digital overlays (patches) and injecting them into camera image messages to manipulate downstream perception. The authors devise three adversarial attack strategies that place overlays of different shapes at user-specified locations, extending prior work that focused on square patches. Overall, task-specific defenses emphasize domain knowledge and application constraints, allowing defenses to be optimized for real-world deployment. However, their limited generality means that techniques developed for one domain may not directly transfer to others.

4.5. Discussion and Insights

The defense methods against adversarial patch attacks demonstrate the diversity of strategies ranging from input-level purification to model-level robustness enhancement. While each category has its unique advantages, several key insights emerge from the comparative analysis:

- **Effectiveness vs. Generality.** Localization and removal-based defenses are intuitive and computationally efficient, but their effectiveness strongly depends on the accuracy of patch detection. In contrast, reconstruction-based approaches using generative models are more general and attack-agnostic, yet they introduce higher computational costs that limit real-time applicability.
- **Integration into Learning Pipelines.** Model modification and training-based defenses are generally more robust, as they directly improve the resilience of the underlying neural network. However, such methods often require retraining or architectural changes, which may not be feasible in scenarios relying on pretrained or large-scale foundation models.
- **Trade-off Between Robustness and Utility.** Most defenses need to carefully balance adversarial robustness with preservation of clean image accuracy and efficiency. Over-aggressive patch suppression or reconstruction may harm benign inputs, reducing the utility of the model in practical settings.

 doi:10.20944/preprints202510.1706.v1

For some earlier defense research (before 2023) [22,24], the defense methods can also be classified as follows: pre-processing methods, model-level defenses, detection-based methods, and certified approaches. Each category provides distinct advantages: pre-processing methods are simple and efficient, model-level defenses offer stronger resilience through training, detection-based methods are suited for safety-critical monitoring, and certified approaches provide formal guarantees. Overall, there is no one-size-fits-all solution to adversarial patch defense.

5. Future Challenges and Research Directions

Although significant progress has been made in understanding and mitigating patch-based adversarial attacks, several critical challenges remain unresolved. In this section, we outline open problems and highlight promising research directions that can guide future efforts toward building robust and trustworthy vision systems.

Generalization to Unseen Attacks. Most existing defenses are tailored to specific attack settings or patch characteristics, such as fixed patch sizes or digital-only environments. However, adversaries can easily adapt strategies by altering patch shapes, textures, or placement, as well as by exploiting novel optimization techniques. Ensuring defense mechanisms that generalize to previously unseen attacks remains a major challenge. Future research should explore meta-learning and adaptive robustness frameworks capable of detecting and mitigating diverse and evolving patch strategies.

Robustness in the Physical World. While digital attacks provide important insights, real-world patch attacks pose a greater threat due to their practicality and accessibility. Defenses that perform well in simulation often degrade significantly under physical conditions, where factors such as lighting, viewpoint variation, printing quality, and occlusions introduce additional complexity. Developing defenses that remain reliable in uncontrolled physical environments is essential, particularly for safety-critical applications such as autonomous driving and biometric authentication.

Scalability to Large-Scale Models and Tasks. As vision systems increasingly rely on large-scale deep learning models, including vision transformers and foundation models, scalability becomes a pressing issue. Many existing defenses are computationally expensive, require retraining, or do not scale effectively to high-resolution inputs and multi-task systems. Future work should focus on designing efficient defenses that maintain robustness without sacrificing scalability, latency, or energy efficiency, enabling deployment in real-world platforms with limited computational resources.

Integration with Broader AI Security Ecosystems. Patch attacks are only one facet of the adversarial threat landscape, which also includes imperceptible perturbations, poisoning, and backdoor attacks. Effective real-world security will likely require unified frameworks that integrate patch defenses with countermeasures for other adversarial risks. Exploring synergistic defense architectures and standardized evaluation benchmarks across attack modalities represents an important step toward comprehensive AI security.

6. Conclusion

Adversarial patch attacks have emerged as a powerful and practical threat to computer vision systems, exploiting localized perturbations to cause severe misbehavior across classification, detection, and other vision-based tasks. In this survey, we provided a comprehensive taxonomy of patch attack methods, highlighting their evolution from digital demonstrations to physical-world implementations and their adaptation to diverse application domains. We further reviewed existing defense strategies, analyzing their strengths, limitations, and suitability for real-world deployment. Despite encouraging progress, defending against adversarial patches remains an open and dynamic research challenge. Current defenses often struggle with generalization to unseen attacks, robustness in the physical world, and scalability to large-scale models and complex tasks. Moreover, trade-offs between robustness and utility, as well as the lack of comprehensive, standardized benchmarks, continue to hinder practical deployment. Addressing these challenges will require the development of adaptive, context-aware, and provably robust methods that integrate seamlessly into broader AI security ecosystems.

 doi:10.20944/preprints202510.1706.v1

This survey will help researchers and practitioners gain a clearer picture of the current landscape of patch attacks and defenses, while also sparking new ideas for future work. Strengthening both the theoretical foundations and practical approaches to patch robustness will bring the community closer to developing computer vision systems that are safe, trustworthy, and resilient against adversarial threats in real-world settings.

Author Contributions: Conceptualization, X.L. and R.X.; methodology, X.L. and R.X.; software, X.L. and R.X.; validation, X.L. and R.X.; formal analysis, X.L. and R.X.; investigation, X.L. and R.X.; resources, X.L. and R.X.; data curation, X.L. and R.X.; writing—original draft preparation, X.L. and R.X.; writing—review and editing, X.L. and R.X.; visualization, X.L. and R.X.; supervision, R.X.; project administration, R.X.; funding acquisition, R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
DL	Deep Learning
DNN	Deep neural networks
GANs	Generative Adversarial Networks
ML	Machine Lerning
ViTs	Vision Transformers

References

1. Ota, K.; Dao, M.S.; Mezaris, V.; Natale, F.G.D. Deep learning for mobile multimedia: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **2017**, *13*, 1–22.
2. Chen, C.; Seff, A.; Kornhauser, A.; Xiao, J. Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 2722–2730.
3. Liu, X.; Xu, R.; Chen, Y. A decentralized digital watermarking framework for secure and auditable video data in smart vehicular networks. *Future Internet* **2024**, *16*.
4. Liu, X.; Xiao, P.; Esposito, M.; Raavi, M.; Zhao, C. AGFA-Net: Attention-Guided Feature-Aggregated Network for Coronary Artery Segmentation Using Computed Tomography Angiography. In Proceedings of the 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2024, pp. 327–334.
5. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)* **2022**, *54*, 1–41.
6. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *International journal of computer vision* **2020**, *128*, 261–318.
7. Chai, J.; Zeng, H.; Li, A.; Ngai, E.W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications* **2021**, *6*, 100134.
8. Guo, Q.; Chen, S.; Xie, X.; Ma, L.; Hu, Q.; Liu, H.; Liu, Y.; Zhao, J.; Li, X. An empirical study towards characterizing deep learning development and deployment across different frameworks and platforms. In Proceedings of the 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2019, pp. 810–822.
9. He, Y.; Meng, G.; Chen, K.; Hu, X.; He, J. Towards security threats of deep learning systems: A survey. *IEEE Transactions on Software Engineering* **2020**, *48*, 1743–1770.
10. Liu, X.; Xu, R.; Peng, X. BEWSAT: blockchain-enabled watermarking for secure authentication and tamper localization in industrial visual inspection. In Proceedings of the Eighth International Conference on Machine Vision and Applications (ICMVA 2025). SPIE, 2025, Vol. 13734, pp. 54–65.

 doi:10.20944/preprints202510.1706.v1

11. Xu, R.; Liu, X.; Nagothu, D.; Qu, Q.; Chen, Y. Detecting Manipulated Digital Entities Through Real-World Anchors. In Proceedings of the International Conference on Advanced Information Networking and Applications. Springer, 2025, pp. 450–461.
12. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.
13. Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* **2019**.
14. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* **2015**.
15. Li, Y.; Xie, B.; Guo, S.; Yang, Y.; Xiao, B. A survey of robustness and safety of 2d and 3d deep learning models against adversarial attacks. *ACM Computing Surveys* **2024**, *56*, 1–37.
16. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems* **2019**, *30*, 2805–2824.
17. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* **2017**.
18. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* **2017**.
19. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665* **2017**.
20. Wei, H.; Tang, H.; Jia, X.; Wang, Z.; Yu, H.; Li, Z.; Satoh, S.; Van Gool, L.; Wang, Z. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 9797–9817.
21. Liu, X.; Shen, F.; Zhao, J.; Nie, C. EAP: An effective black-box impersonation adversarial patch attack method on face recognition in the physical world. *Neurocomputing* **2024**, *580*, 127517.
22. Sharma, A.; Bian, Y.; Munz, P.; Narayan, A. Adversarial patch attacks and defences in vision-based tasks: A survey. *arXiv preprint arXiv:2206.08304* **2022**.
23. Wang, D.; Yao, W.; Jiang, T.; Tang, G.; Chen, X. A survey on physical adversarial attack in computer vision. *arXiv preprint arXiv:2209.14262* **2022**.
24. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology* **2021**, *6*, 25–45.
25. Liu, X.; Liu, Z.; Chatterjee, S.; Portfleet, M.; Sun, Y. Understanding human behaviors and injury factors in underground mines using data analytics. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021, pp. 2459–2462.
26. Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; Tao, D. Perceptual-sensitive gan for generating adversarial patches. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 1028–1035.
27. Kang, H.; Kim, H.; et al. Robust adversarial attack against explainable deep classification models based on adversarial images with different patch sizes and perturbation ratios. *IEEE Access* **2021**, *9*, 133049–133061.
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*.
29. Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **2017**, *29*, 2352–2449.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
32. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
33. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.
34. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

 doi:10.20944/preprints202510.1706.v1

35. Liu, X.; Yang, H.; Liu, Z.; Song, L.; Li, H.; Chen, Y. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299* **2018**.
36. Song, D.; Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Tramer, F.; Prakash, A.; Kohno, T. Physical adversarial examples for object detectors. In Proceedings of the 12th USENIX workshop on offensive technologies (WOOT 18), 2018.
37. Yamanaka, K.; Matsumoto, R.; Takahashi, K.; Fujii, T. Adversarial patch attacks on monocular depth estimation networks. *IEEE Access* **2020**, *8*, 179094–179104.
38. Mirsky, Y. Ipatch: A remote adversarial patch. *Cybersecurity* **2023**, *6*, 18.
39. Ran, Y.; Wang, W.; Li, M.; Li, L.C.; Wang, Y.G.; Li, J. Cross-shaped adversarial patch attack. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**, *34*, 2289–2303.
40. Tao, G.; An, S.; Cheng, S.; Shen, G.; Zhang, X. Hard-label black-box universal adversarial patch attack. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 697–714.
41. Li, C.; Yan, H.; Zhou, L.; Chen, T.; Liu, Z.; Su, H. Prompt-guided environmentally consistent adversarial patch. *arXiv preprint arXiv:2411.10498* **2024**.
42. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* **2014**.
43. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models. In Proceedings of the Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec), 2017, pp. 15–26.
44. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box adversarial attacks with limited queries and information. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 2137–2146.
45. Bhagoji, A.N.; He, W.; Li, B.; Song, D. Practical black-box attacks on deep neural networks using efficient query mechanisms. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 158–174.
46. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* **2013**.
47. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017, pp. 506–519.
48. Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M.K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the Proceedings of the 2016 acm sigsac conference on computer and communications security, 2016, pp. 1528–1540.
49. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1625–1634.
50. Karmon, D.; Zoran, D.; Goldberg, Y. Lavan: Localized and visible adversarial noise. In Proceedings of the International conference on machine learning. PMLR, 2018, pp. 2507–2515.
51. Thys, S.; Van Ranst, W.; Goedemé, T. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detectors. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 0–7.
52. Hingun, N.; Sitawarin, C.; Li, J.; Wagner, D. REAP: A Large-Scale Realistic Adversarial Patch Benchmark. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12500–12509.
53. Shrestha, S.; Pathak, S.; Viegas, E.K. Towards a robust adversarial patch attack against unmanned aerial vehicles object detection. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023, pp. 3256–3263.
54. Wang, X.; Li, W. Physical adversarial attacks for infrared object detection. In Proceedings of the 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2024, pp. 64–69.
55. Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; Lyu, S. Invisible backdoor attack with sample-specific triggers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 16463–16472.
56. Zhou, X.; Pan, Z.; Duan, Y.; Zhang, J.; Wang, S. A data independent approach to generate adversarial patches. *Machine Vision and Applications* **2021**, *32*, 67.

 doi:10.20944/preprints202510.1706.v1

57. Wang, Z.; Huang, J.J.; Liu, T.; Chen, Z.; Zhao, W.; Liu, X.; Pan, Y.; Liu, L. Multi-patch adversarial attack for remote sensing image classification. In Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Springer, 2023, pp. 377–391.
58. Huang, J.J.; Wang, Z.; Liu, T.; Luo, W.; Chen, Z.; Zhao, W.; Wang, M. DeMPAA: Deployable multi-mini-patch adversarial attack for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–13.
59. Tiliwalidi, K.; Yan, K.; Shi, Y.; Hu, C.; Zhou, J. Cost-effective and robust adversarial patch attacks in real-world scenarios. *Journal of Electronic Imaging* **2025**, *34*, 033003–033003.
60. Lee, M.; Kolter, Z. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897* **2019**.
61. Wu, S.; Dai, T.; Xia, S.T. Dpattack: Diffused patch attacks against universal object detection. *arXiv preprint arXiv:2010.11679* **2020**.
62. Huang, H.; Wang, Y.; Chen, Z.; Tang, Z.; Zhang, W.; Ma, K.K. Rpattack: Refined patch attack on general object detectors. *arXiv preprint arXiv:2103.12469* **2021**.
63. Hoory, S.; Shapira, T.; Shabtai, A.; Elovici, Y. Dynamic adversarial patch for evading object detection models. *arXiv preprint arXiv:2010.13070* **2020**.
64. Wang, Y.; Lv, H.; Kuang, X.; Zhao, G.; Tan, Y.a.; Zhang, Q.; Hu, J. Towards a physical-world adversarial patch for blinding object detection models. *Information Sciences* **2021**, *556*, 459–471.
65. Lang, D.; Chen, D.; Shi, R.; He, Y. Attention-Guided Digital Adversarial Patches on Visual Detection. *Security and Communication Networks* **2021**, *2021*, 6637936.
66. Sun, X.; Cheng, G.; Pei, L.; Li, H.; Han, J. Threatening patch attacks on object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–10.
67. Tang, G.; Jiang, T.; Zhou, W.; Li, C.; Yao, W.; Zhao, Y. Adversarial patch attacks against aerial imagery object detectors. *Neurocomputing* **2023**, *537*, 128–140.
68. Deng, B.; Zhang, D.; Dong, F.; Zhang, J.; Shafiq, M.; Gu, Z. Rust-style patch: A physical and naturalistic camouflage attacks on object detector for remote sensing images. *Remote Sensing* **2023**, *15*, 885.
69. Wang, J.; Li, F.; He, L. A unified framework for adversarial patch attacks against visual 3D object detection in autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology* **2025**.
70. Zhu, W.; Ji, X.; Cheng, Y.; Zhang, S.; Xu, W. {TPatch}: A triggered physical adversarial patch. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 661–678.
71. Wei, H.; Wang, Z.; Zhang, K.; Hou, J.; Liu, Y.; Tang, H.; Wang, Z. Revisiting adversarial patches for designing camera-agnostic attacks against person detection. *Advances in Neural Information Processing Systems* **2024**, *37*, 8047–8064.
72. Zhang, S.; Cheng, Y.; Zhu, W.; Ji, X.; Xu, W. {CAPatch}: Physical Adversarial Patch against Image Captioning Systems. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 679–696.
73. Jiang, K.; Chen, Z.; Huang, H.; Wang, J.; Yang, D.; Li, B.; Wang, Y.; Zhang, W. Efficient decision-based black-box patch attacks on video recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4379–4389.
74. Liu, A.; Guo, J.; Wang, J.; Liang, S.; Tao, R.; Zhou, W.; Liu, C.; Liu, X.; Tao, D. {X-Adv}: Physical adversarial object attacks against x-ray prohibited item detection. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 3781–3798.
75. Agrawal, K.; Bhatnagar, C. A black-box based attack generation approach to create the transferable patch attack. In Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2023, pp. 1376–1380.
76. Wei, X.; Yu, J.; Huang, Y. Infrared adversarial patches with learnable shapes and locations in the physical world. *International Journal of Computer Vision* **2024**, *132*, 1928–1944.
77. Chen, X.; Liu, F.; Jiang, D.; Yan, K. Natural adversarial patch generation method based on latent diffusion model. *arXiv preprint arXiv:2312.16401* **2023**.
78. Lin, C.Y.; Huang, T.Y.; Ng, H.F.; Lin, W.Y.; Farady, I. Entropy-Boosted Adversarial Patch for Concealing Pedestrians in YOLO Models. *IEEE Access* **2024**, *12*, 32772–32779.
79. Zhou, Z.; Zhao, H.; Liu, J.; Zhang, Q.; Geng, L.; Lyu, S.; Feng, W. Mvpatch: More vivid patch for adversarial camouflaged attacks on object detectors in the physical world. *arXiv preprint arXiv:2312.17431* **2023**.
80. Wang, S.; Li, W.; Xu, Z.; Yu, N. Learning from the Environment: A Novel Adversarial Patch Attack against Object Detectors Using a GAN Trained on Image Slices. In Proceedings of the 2025 2nd International Conference on Electronic Engineering and Information Systems (EEISS). IEEE, 2025, pp. 1–6.

 doi:10.20944/preprints202510.1706.v1

81. Zhou, D.; Qu, H.; Wang, N.; Peng, C.; Ma, Z.; Yang, X.; Gao, X. Fooling human detectors via robust and visually natural adversarial patches. *Neurocomputing* **2025**, *616*, 128915.
82. Yuan, S.; Li, H.; Han, X.; Xu, G.; Jiang, W.; Ni, T.; Zhao, Q.; Fang, Y. Itpatch: An invisible and triggered physical adversarial patch against traffic sign recognition. *arXiv preprint arXiv:2409.12394* **2024**.
83. Hu, Z.; Yang, X.; Zhao, J.; Gao, H.; Xu, H.; Mu, H.; Wang, Y. Physically structured adversarial patch inspired by natural leaves multiply angles deceives infrared detectors. *Journal of King Saud University-Computer and Information Sciences* **2024**, *36*, 102122.
84. Liu, T.; Yang, C.; Liu, X.; Han, R.; Ma, J. RPAU: Fooling the eyes of UAVs via physical adversarial patches. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *25*, 2586–2598.
85. Nesti, F.; Rossolini, G.; Nair, S.; Biondi, A.; Buttazzo, G. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2280–2289.
86. Chen, J.; Zhang, Y.; Liu, C.; Chen, K.; Zou, Z.; Shi, Z. Digital-to-Physical visual consistency optimization for adversarial patch generation in remote sensing scenes. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–17.
87. Wiyatno, R.R.; Xu, A. Physical adversarial textures that fool visual object tracking. In Proceedings of the Proceedings of the IEEE/CVF International Conference on computer vision, 2019, pp. 4822–4831.
88. Ranjan, A.; Janai, J.; Geiger, A.; Black, M.J. Attacking optical flow. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2404–2413.
89. Liu, S.; Wang, J.; Liu, A.; Li, Y.; Gao, Y.; Liu, X.; Tao, D. Harnessing perceptual adversarial patches for crowd counting. In Proceedings of the Proceedings of the 2022 ACM SIGSAC conference on computer and communications security, 2022, pp. 2055–2069.
90. Tarchoun, B.; Ben Khalifa, A.; Mahjoub, M.A.; Abu-Ghazaleh, N.; Alouani, I. Jedi: Entropy-based localization and removal of adversarial patches. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 4087–4095.
91. Xu, K.; Xiao, Y.; Zheng, Z.; Cai, K.; Nevatia, R. Patchzero: Defending against adversarial patch attacks by detecting and zeroing the patch. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4632–4641.
92. Xiang, C.; Valtchanov, A.; Mahloujifar, S.; Mittal, P. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023, pp. 1329–1347.
93. Jing, L.; Wang, R.; Ren, W.; Dong, X.; Zou, C. PAD: Patch-agnostic defense against adversarial patch attacks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24472–24481.
94. Bunzel, N.; Siwakoti, A.; Klause, G. Adversarial patch detection and mitigation by detecting high entropy regions. In Proceedings of the 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2023, pp. 124–128.
95. Hofman, O.; Giloni, A.; Hayun, Y.; Morikawa, I.; Shimizu, T.; Elovici, Y.; Shabtai, A. X-detect: Explainable adversarial patch detection for object detectors in retail. *Machine Learning* **2024**, *113*, 6273–6292.
96. Wu, S.; Wang, J.; Zhao, J.; Wang, Y.; Liu, X. NAPGuard: Towards detecting naturalistic adversarial patches. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24367–24376.
97. Victorica, M.B.; Dán, G.; Sandberg, H. Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 20360–20369.
98. Gibert, D.; Zizzo, G.; Le, Q. Certified robustness of static deep learning-based malware detectors against patch and append attacks. In Proceedings of the Proceedings of the 16th ACM workshop on artificial intelligence and security, 2023, pp. 173–184.
99. Kang, C.; Dong, Y.; Wang, Z.; Ruan, S.; Chen, Y.; Su, H.; Wei, X. Diffender: Diffusion-based adversarial defense against patch attacks. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 130–147.
100. Lin, Z.; Zhao, Y.; Chen, K.; He, J. I don't know you, but I can catch you: Real-time defense against diverse adversarial patches for object detectors. In Proceedings of the Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024, pp. 3823–3837.

101. Liu, X.; Shen, F.; Zhao, J.; Nie, C. Radap: A robust and adaptive defense against diverse adversarial patches on face recognition. *Pattern Recognition* **2025**, *157*, 110915.
102. Wei, X.; Kang, C.; Dong, Y.; Wang, Z.; Ruan, S.; Chen, Y.; Su, H. Real-world adversarial defense against patch attacks based on diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
103. Cai, J.; Chen, S.; Li, H.; Xia, B.; Mao, Z.; Yuan, W. HARP: Let object detector undergo hyperplasia to counter adversarial patches. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 2673–2683.
104. Chen, Z.; Dash, P.; Pattabiraman, K. Jujutsu: A two-stage defense against adversarial patch attacks on deep neural networks. In Proceedings of the Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, 2023, pp. 689–703.
105. Yu, C.; Chen, J.; Wang, Y.; Xue, Y.; Ma, H. Improving adversarial robustness against universal patch attacks through feature norm suppressing. *IEEE Transactions on Neural Networks and Learning Systems* **2023**.
106. Zheng, Y.; Demetrio, L.; Cinà, A.E.; Feng, X.; Xia, Z.; Jiang, X.; Demontis, A.; Biggio, B.; Roli, F. Hardening RGB-D object recognition systems against adversarial patch attacks. *Information Sciences* **2023**, *651*, 119701.
107. Chattopadhyay, N.; Guesmi, A.; Hanif, M.A.; Ouni, B.; Shafique, M. Oddr: Outlier detection & dimension reduction based defense against adversarial patches. *arXiv preprint arXiv:2311.12084* **2023**.
108. Liang, J.; Yi, R.; Chen, J.; Nie, Y.; Zhang, H. Securing autonomous vehicles visual perception: Adversarial patch attack and defense schemes with experimental validations. *IEEE Transactions on Intelligent Vehicles* **2024**.
109. Chattopadhyay, N.; Guesmi, A.; Shafique, M. Anomaly unveiled: Securing image classification against adversarial patch attacks. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024, pp. 929–935.
110. Strack, L.; Waseda, F.; Nguyen, H.H.; Zheng, Y.; Echizen, I. Defending against physical adversarial patch attacks on infrared human detection. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024, pp. 3896–3902.
111. Wu, H.; Yunas, S.; Rowlands, S.; Ruan, W.; Wahlström, J. Adversarial detection: Attacking object detection in real time. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2023, pp. 1–7.