Beyond Vulnerabilities: A Survey of Adversarial Attacks as Both Threats and Defenses in Computer Vision Systems

Zhongliang Guo **Yifei Qian** **Yanli Li** **Weiye Li** **Chun Tong Lei** **Shuai Zhao** **Lei Fang**

Ognjen Arandjelović **Chun Pong Lau**

Abstract

Adversarial attacks against computer vision systems have emerged as a critical research area that challenges the fundamental assumptions about neural network robustness and security. This comprehensive survey examines the evolving landscape of adversarial techniques, revealing their dual nature as both sophisticated security threats and valuable defensive tools. We provide a systematic analysis of adversarial attack methodologies across three primary domains: pixel-space attacks, physically realizable attacks, and latent-space attacks. Our investigation traces the technical evolution from early gradient-based methods such as FGSM and PGD to sophisticated optimization techniques incorporating momentum, adaptive step sizes, and advanced transferability mechanisms. We examine how physically realizable attacks have successfully bridged the gap between digital vulnerabilities and real-world threats through adversarial patches, 3D textures, and dynamic optical perturbations. Additionally, we explore the emergence of latent-space attacks that leverage semantic structure in internal representations to create more transferable and meaningful adversarial examples. Beyond traditional offensive applications, we investigate the constructive use of adversarial techniques for vulnerability assessment in biometric authentication systems and protection against malicious generative models. Our analysis reveals critical research gaps, particularly in neural style transfer protection and computational efficiency requirements. This survey contributes a comprehensive taxonomy, evolution analysis, and identification of future research directions, aiming to advance understanding of adversarial vulnerabilities and inform the development of more robust and trustworthy computer vision systems.

1 Introduction

The remarkable advancement of deep neural networks (DNNs) has revolutionized artificial intelligence applications across numerous domains, with computer vision systems achieving unprecedented performance levels that often match or exceed human capabilities [1, 2, 3]. This technological evolution has led to widespread deployment of vision-based AI systems in critical sectors including autonomous driving, medical diagnosis, financial services, and security applications. However, alongside these achievements, fundamental vulnerabilities have emerged that challenge the reliability and trustworthiness of these systems.

A pivotal discovery by Szegedy et al. [4] revealed that neural networks are susceptible to adversarial attacks—carefully crafted inputs that can cause models to produce incorrect outputs while remaining imperceptible to human observers, as shown in Figure 1. This vulnerability represents a significant security concern, as minor pixel-level perturbations can systematically fool state-of-the-art computer vision models. The implications extend far beyond academic curiosity, manifesting as real-world threats in safety-critical applications where model failures can have serious consequences.

Simultaneously, the rapid development of generative AI technologies has introduced new ethical and security challenges. The emergence of sophisticated generative models capable of producing highly realistic synthetic content has enabled both beneficial applications and malicious uses, including deepfakes [7], unauthorized content generation [8, 9], and intellectual property violations [10, 11]. This dual nature of AI technology—as both a powerful tool and a

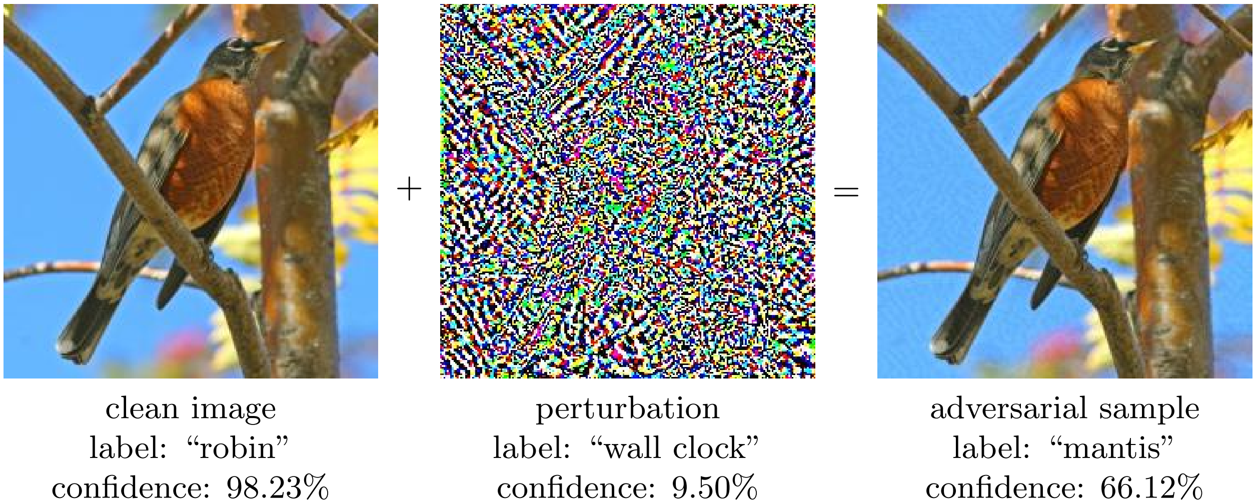


Figure 1: An exemplar for a successful adversarial attack. The attack is implemented by the Fast Gradient Sign Method [5] on VGG-19 [6] with eps=4/255. The clean image is correctly predicted by a classification model as robin with 98.23% confidence. However, when We add perturbations that is not up to 4/255 for each pixel, although perturbations are imperceptible for human, the classification model predicts the adversarial sample as a wrong category.

potential security risk—underscores the critical need for comprehensive understanding of adversarial vulnerabilities and protection mechanisms.

The implications of these vulnerabilities extend far beyond academic concerns, manifesting across critical real-world applications. In autonomous driving systems, adversarial patches on traffic signs can cause misclassification, potentially leading to safety hazards [12]. Medical imaging systems face similar risks, where imperceptible perturbations to X-rays or CT scans could result in misdiagnosis [13]. Financial institutions employing ML for fraud detection are vulnerable to adversarial examples that could allow fraudulent transactions to bypass security measures [14]. Content moderation systems on social media platforms can be fooled by adversarial examples, enabling the dissemination of harmful content [15]. Industrial quality control systems represent another critical vulnerability, where adversarial attacks could cause defective products to pass visual inspection [16]. Conversely, adversarial techniques have shown promise in beneficial applications, such as protecting individual privacy from unauthorized facial recognition systems [17] and safeguarding intellectual property of trained models from unauthorized extraction [18]. These diverse applications underscore the critical importance of understanding and addressing adversarial vulnerabilities in modern ML systems.

1.1 Research Scope and Objectives

This literature review provides a comprehensive examination of adversarial attacks in computer vision systems, focusing on three primary research dimensions:

1. **Attack Methodologies:** A systematic analysis of adversarial attack techniques, including their underlying mechanisms, effectiveness, and evolution from basic perturbation methods to sophisticated attack strategies.
2. **Vulnerability Assessment:** An exploration of how adversarial techniques serve as tools for evaluating the robustness boundaries of computer vision models, particularly in critical applications such as biometric authentication systems.
3. **Defensive Applications:** An investigation into the constructive use of adversarial techniques for preventing malicious applications, including protection against unauthorized neural style transfer and misuse of generative models.

The scope of this review encompasses pixel-space attacks, physically realizable attacks, and latent-space attacks, while also examining emerging applications in biometric security and generative model protection. We aim to provide researchers and practitioners with a comprehensive understanding of the current state of adversarial machine learning in computer vision and identify key research gaps and future directions.

1.2 Contribution and Organization

This review makes several key contributions to the field:

- **Comprehensive Taxonomy:** We present a systematic categorization of adversarial attacks based on attacker knowledge, underlying mechanisms, objectives, and operational domains.
- **Evolution Analysis:** We trace the technical evolution from early gradient-based methods to sophisticated optimization techniques and their practical implications.
- **Application-Focused Perspective:** We examine adversarial techniques from both offensive and defensive standpoints, highlighting their dual role in vulnerability assessment and protection mechanisms.
- **Critical Gap Identification:** We identify underexplored research areas and emerging challenges that warrant future investigation.

The remainder of this review is organized as follows: Section 2 establishes fundamental concepts and evaluation frameworks for adversarial attacks. Sections 3-5 provide detailed examinations of pixel-space attacks, physically realizable attacks, and latent-space attacks respectively. Section 6 explores emerging applications in biometric security and generative model protection. Finally, Section 7 discusses future research directions and concludes the review.

Through this comprehensive analysis, we aim to advance understanding of adversarial vulnerabilities in computer vision systems and inform the development of more robust and trustworthy AI technologies.

2 Preliminary

This section provides an elementary introduction to adversarial attacks in computer vision systems. We begin by exploring the fundamental concepts of adversarial attacks, presenting a systematic categorization based on attacker knowledge, underlying mechanisms, objectives, and domains of operation. We also discuss the evaluation metrics used to assess the effectiveness of these attacks, including performance, transferability, and imperceptibility.

2.1 Concept of Adversarial Attacks

This paper start with the concept of adversarial attack. Given an image $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ from a dataset $\mathcal{X}$, where C, H, W is the number of color channels, the image height, and its width, respectively. An adversarial example $\boldsymbol{x}^{adv}$ is generated as follows:

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \boldsymbol{\delta}, \tag{1}$$

where $\boldsymbol{\delta}$ refers to the adversarial perturbation. Here We unify any attack method as $A(\cdot)$, having $\boldsymbol{\delta} = A(\boldsymbol{x})$. The NN, denoted as $\mathcal{F}(\cdot)$, serves as a differentiable model that, when given an input, produces an output reflective of the specific task it is designed to perform. The adversarial perturbations will significantly deviate the output of a NN from its expected response. If we leverage $\mathcal{D}(\cdot)$ to measure the difference between pre-attack results and post-attack results, it can be denoted as $\max(\mathcal{D}(\mathcal{F}(\boldsymbol{x}), \mathcal{F}(\boldsymbol{x}^{adv})))$.

In the majority of attack scenarios, adversarial perturbations are confined within a specific range ϵ to ensure they remain visually imperceptible [5, 19]. The measure of range is various, may include quantitatively controlling the size of the perturbation, or controlling the human perception changes towards the image. This constraint allows the perturbations to subtly manipulate the NN's output without raising suspicion or being detectable by human observers.

Thus, an adversarial attack method can be formulated as:

$$\arg\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \mathcal{D}\big(\mathcal{F}(\boldsymbol{x}), \mathcal{F}(\boldsymbol{x}^{adv})\big)\,, \tag{2}$$

where $\|\cdot\|$ denotes a kind of constraint, to maintain the visual integrity of adversarial samples. In the practical execution of adversarial attacks, The constraint $\|\cdot\|$ can represent various forms of limitations, such as the maximum allowable perturbation on individual pixels [5], ensuring minimal visual perceptibility [20], or maintaining the semantic content of the image [21]. This formulation encapsulates the essence of crafting adversarial examples that are effective in deceiving the model while being subtle enough to evade detection.

2.2 Categories of Adversarial Attacks

Adversarial attacks can be systematically categorized according to several fundamental criteria [22, 23]. In this section, We will elucidate the mainstream taxonomies used to classify these attacks, including the attacker's knowledge of the target model, the underlying mechanisms employed, the specific objectives pursued, and the domains in which these attacks operate.

2.2.1 Knowledge of Attackers

Regarding to the attacker's knowledge about the target model, i.e., if the attacker can access the NN architecture and weights, the terms "black-box" and "white-box" refer to two distinct scenarios.

White-box attack. White-box attack is one where the attacker possesses complete knowledge of the NN's architecture and parameters. This comprehensive insight allows the attacker to craft adversarial examples with precision, exploiting the specific vulnerabilities of the model. The attacker can utilize gradient-based methods [5, 19] or other optimization techniques [24, 25] to generate inputs that are designed to lead the model astray, making white-box attacks particularly potent and challenging to defend against.

Grey-box attack. Grey-box attack represents an intermediate scenario where the attacker possesses partial knowledge about the target model, such as the architecture but not the parameters, or access to similar training data without knowing the exact model weights. This limited information allows attackers to craft more effective adversarial examples compared to pure black-box scenarios by leveraging architectural similarities or training surrogate models with known components [26, 27]. While less powerful than white-box attacks, grey-box approaches can significantly improve attack success rates by exploiting available structural knowledge about the target model.

Black-box attack. Black-box attack depicts a situation where the attacker has no direct access to the model's internals. The attacker's knowledge is limited to the input data and the corresponding outputs from the model. Despite this lack of detailed information, attackers can still craft effective adversarial examples. They often employ techniques like transferability of adversarial examples from substitute models [28], query-based methods [29], or gradient estimation strategies [30].

2.2.2 Underlying Mechanisms

Adversarial attack techniques can be categorized based on their underlying mechanisms, broadly falling into two distinct types: gradient-based or optimization-based, and generative model-based approaches. Each of these offers different advantages and operates under varying assumptions about the attacker's knowledge and capabilities, contributing to the diverse and complex landscape of adversarial attacks in NNs.

Gradient-based (optimization-based) method. These methods leverage the gradient of the NN to identify the most effective alterations to the input data [5]. Essentially, this kind of approach is to optimize a single image or adversarial perturbation to make the final output deceptive. By understanding how small changes in the input can significantly impact the output, these methods efficiently generate adversarial examples that are tailored to disrupt the network.

Generative model-based method. Generative model-based attack employs generative models like Generative Adversarial Networks (GANs), Autoencoders (AEs), and Diffusion Models (DMs) to craft adversarial examples [31]. These models are trained to generate inputs that are perceptually indistinguishable from real data but are structured in a way that leads the victim network to make incorrect predictions. The essence of this method is to use NNs to learn a distribution in which the images in this distribution are deceptive to the victim model.

2.2.3 Attack Objectives

In the exploration of adversarial strategies, a crucial distinction lies in the objectives of the attacks, specifically differentiating between targeted and untargeted attacks. This differentiation is not limited to classification models or solely to the realm of computer vision but extends across various domains and tasks within computer vision.

Targeted Attacks. the adversary's goal of targeted attacks is to manipulate the model into producing a specific, incorrect output. The attacker aims not just to cause a misclassification or error but to steer the model toward a predetermined, incorrect result. This type of attack requires a more nuanced understanding of the model's behavior and often involves sophisticated perturbation techniques. In contexts beyond classification, such as object detection or segmentation in computer vision, a targeted attack might aim to make the model identify an object as a particular class erroneously or modify the boundaries of segmentation to match a specific, incorrect shape.

Untargeted Attacks. Contrastingly, untargeted attacks focus on the broader objective of simply causing the model to err, without any preference for the specific type of mistake. The primary aim is to degrade the overall performance of the model, ensuring that the output deviates from the correct or expected result, regardless of the specific incorrect outcome. In broader applications, such as in natural language processing or generative models, an untargeted attack could aim to disrupt the coherence of generated text or the fidelity of synthesized images without guiding the model toward any particular error.

2.2.4 Attack Domain

Adversarial attacks can be categorized based on the domain in which they operate, reflecting different approaches to manipulating inputs and exploiting model vulnerabilities.

Pixel-Space Attacks. Pixel-Space Attacks (PSAs) directly modify the raw input images at the pixel level to generate adversarial examples [5, 19]. These attacks operate on the most immediate and explicit representation of visual data, making precise alterations to individual pixel values to induce misclassification. While the modifications are often imperceptible to human observers, they can significantly impact model predictions by exploiting the sensitivity of neural networks to small input changes. PSAs typically leverage the gradient-based / optimization-based framework, though there are few methods [32, 31] used generative models to generate the pixel-space perturbations. PSAs often use ℓ_p-norms as constrains respect to perturbations, ensuring the imperceptibility of adversarial samples.

Physically Realizable Attacks. Physically realizable attacks bridge the gap between digital perturbations and real-world implementations, addressing the challenges of creating adversarial examples that remain effective when captured through sensors or cameras [12]. These attacks consider physical constraints such as lighting conditions, viewing angles, printing limitations, and environmental factors. Examples include adversarial patches, which can be printed and attached to objects; adversarial eyeglass frames that fool facial recognition systems; or adversarial textures applied to 3D objects. The defining characteristic of these attacks is their robustness against physical transformations and their ability to impact systems in practical deployment scenarios.

Latent-Space Attacks. Latent-Space Attacks (LSAs) operate by manipulating the intermediate representations or embeddings that deep learning models construct internally [21]. Rather than modifying raw inputs, these attacks target the higher-dimensional feature spaces where semantic information is encoded. By perturbing these latent representations, attackers can achieve more efficient and semantically meaningful adversarial examples, often with greater transferability across different models. Such attacks leverage the understanding that models organize information in their hidden layers according to learned patterns and concepts, making these internal representations particularly vulnerable.

2.3 Evaluation Protocol for Adversarial Attacks

The evaluation protocol for adversarial attack can be divided into three aspects, performance, transferability, and imperceptibility. The challenge in adversarial attack design lies in balancing attack effectiveness with imperceptibility, as these goals often conflict: stronger perturbations typically increase attack success rates but become more visible to human observers.

2.3.1 Performance

The efficacy of adversarial attacks is evaluated using different metrics, depending on the type of model being targeted. These metrics quantify the success rate of perturbations in causing model errors.

Classification Models. For classification models, the Attack Success Rate (ASR) serves as the primary metric, defined as the percentage of adversarial examples that successfully cause misclassification. For targeted attacks, ASR measures the percentage of samples that are misclassified as the attacker's chosen target class. For untargeted attacks, it measures the percentage of samples that are simply misclassified (regardless of which incorrect class is chosen). Mathematically, ASR can be expressed as:

$$\text{ASR} = \frac{\text{Number of successful adversarial examples}}{\text{Total number of adversarial examples}} \times 100\%. \tag{3}$$

Higher ASR values indicate more effective attacks. When evaluating defense mechanisms, a lower ASR demonstrates greater robustness against adversarial perturbations.

Object Detection Models. For object detection models, attack performance is evaluated using metrics that address both localization and classification aspects. Common metrics include:

- **Mean Average Precision (mAP) drop**: the reduction in mAP between clean and adversarial inputs.
- **Detection rate reduction**: The percentage decrease in correctly detected objects.

- **False positive increase**: The percentage increase in falsely detected objects.
- **Disappearance rate**: The percentage of objects that are completely missed by the detector after perturbation.

These metrics capture different aspects of disruption to object detection systems, with successful attacks typically causing significant degradation across multiple metrics.

Segmentation Models. For segmentation models, which predict pixel-wise class labels, attack performance is evaluated using segmentation-specific metrics:

- **Intersection over Union (IoU) reduction**: The decrease in overlap between ground truth and predicted segmentation masks.
- **Mean IoU (mIoU) drop**: The average IoU reduction across all classes.
- Pixel accuracy decrease: The reduction in correctly classified pixels.
- **Boundary F1 score reduction**: The decrease in accuracy of object boundaries.

Effective attacks against segmentation models typically result in substantial decreases in these metrics, indicating significantly degraded segmentation quality.

2.3.2 Transferability

Transferability measures the effectiveness of adversarial examples when applied to models different from the one used to generate them. This property is particularly important for black-box attacks, where direct access to the target model is limited or unavailable. Transferability is typically evaluated by:

- **Cross-model transferability**: Success rate when transferring between different model architectures (e.g., from ResNet [33] to VGG [6]).
- **Cross-task transferability**: Success rate when transferring between different tasks (e.g., from classification to object detection).
- **Cross-dataset transferability**: ASR when transferring between models trained on different datasets.

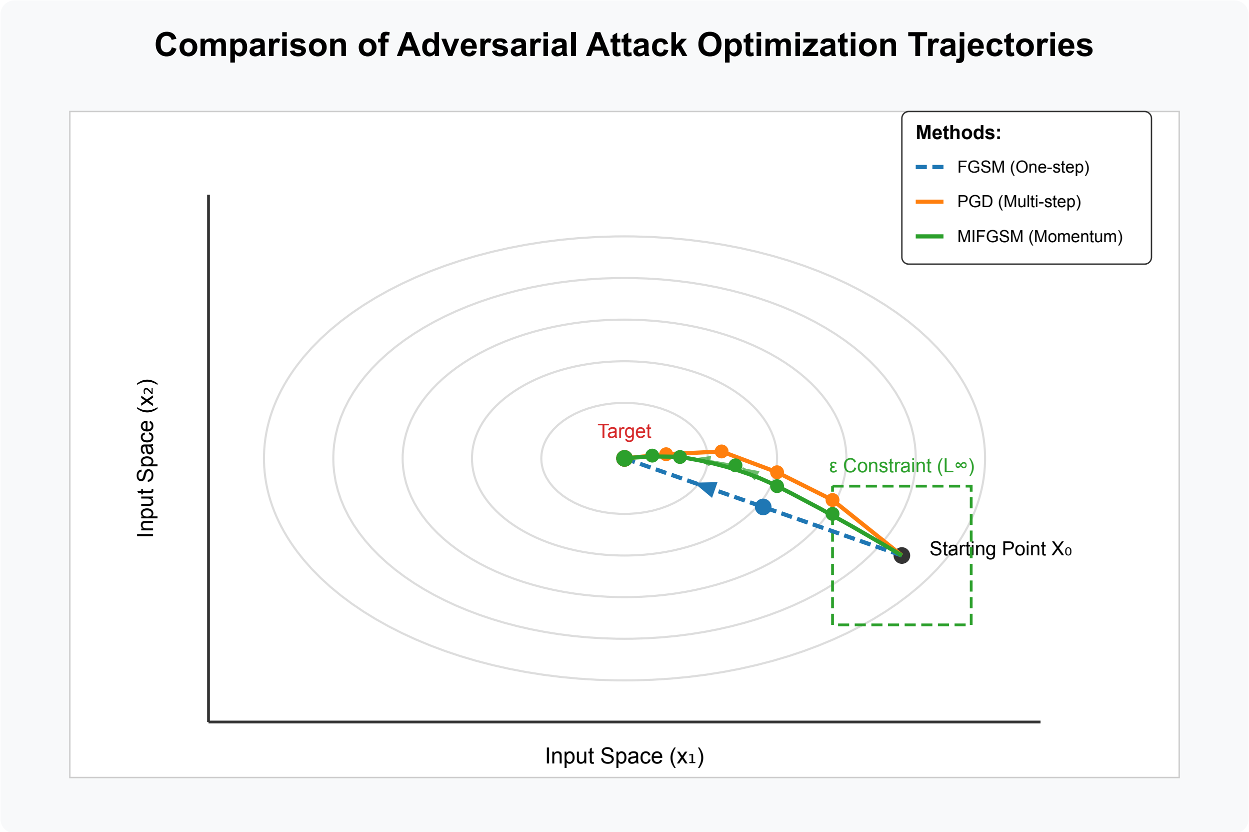
Highly transferable attacks are generally considered more powerful, as they indicate the exploitation of fundamental vulnerabilities rather than model-specific weaknesses. Transferability is quantified by calculating the ASR or performance degradation metrics on the transfer model compared to the source model.

2.3.3 Imperceptibility

Imperceptibility measures how difficult it is for humans to detect the adversarial perturbations, which is crucial for practical attacks that must remain unnoticed. Various Image Quality Assessments (IQAs) are used to evaluate imperceptibility:

- **Peak Signal-to-Noise Ratio (PSNR)**: Measures the ratio between the maximum possible power of a signal and the power of corrupting noise. Higher PSNR values indicate greater similarity between the original and perturbed images.
- **Structural Similarity Index Measure (SSIM)**: Evaluates the similarity of two images based on luminance, contrast, and structure, with values ranging from -1 to 1 [34]. SSIM values closer to 1 indicate greater similarity.
- **Fréchet Inception Distance (FID)**: Measures the distance between the feature representations of real and generated images [35]. Lower FID scores indicate greater perceptual similarity.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Utilizes deep neural networks to measure perceptual similarity, better aligning with human visual perception than traditional metrics [36]. Lower LPIPS values indicate greater perceptual similarity.

Additionally, human perception studies are sometimes conducted to evaluate imperceptibility directly, as computational metrics do not always perfectly align with human visual perception. In these studies, participants are asked to distinguish between original and adversarial images, with higher success rates indicating less imperceptible perturbations.



Method	Iterations	Key Features	Update Formula
FGSM	Single-step	Fast but limited effectiveness Direct gradient approach	$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y))$
PGD	Multi-step	Stronger attack with projection Can follow complex loss surfaces	$x_{t+1}^{adv} = \Pi(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x J(x_t^{adv}, y)))$
MI-FGSM		Better transferability Avoids local noise & oscillation	$g_{t+1} = \mu \cdot g_t + \nabla_x J(x_t^{adv}, y) \Vert \nabla_x J(x_t^{adv}, y) \Vert_1$ $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})$

Figure 2: Visual optimization comparison of FGSM, PGD, and MI-FGSM.

3 Pixel-Space Attacks

Pixel-space attacks (PSAs) involve modifying pixel values within certain constraints to generate adversarial samples that can deceive machine learning models. These attacks represent the earliest category of adversarial methods proposed in the literature and continue to receive the most extensive attention from researchers. As this field has evolved, significant advancements have been made in improving both the transferability of these attacks across different models and their imperceptibility to human observers. This section provides a comprehensive review of the developments in PSA, mainly focusing on gradient-based attacks, examining their methodologies and the progressive improvements in their effectiveness and sophistication.

3.1 Early Foundations and Momentum-Based Enhancements

As shown in Figure 2, the Fast Gradient Sign Method (FGSM) [5] and its iterative variants (I-FGSM [37], PGD [19]) represent the first generation of PSA. These methods employ the gradient of the given loss function with respect to the input image to determine the update of perturbation. Specifically, FGSM generates adversarial examples x^{adv} from clean images x by one step:

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{J}(x, y)), \tag{4}$$

where $\nabla_x \mathcal{J}(x, y)$ is the gradient of the loss function $\mathcal{J}$ with respect to input x for true label y, and ϵ is the perturbation magnitude.

A significant progress was made by the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [24]. This approach incorporated momentum terms into the iterative attack optimization process, substantially improving transferability across different victim models. The update process of MI-FGSM is:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y)}{||\nabla_x J(x_t^{adv}, y)||_1}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1}), \tag{5}$$

where μ is the momentum decay factor, g_t is the accumulated gradient at iteration t, and α is the step size, or generally speaking, learning rate.

MI-FGSM's high performance laid the foundation for further momentum-based enhancements. Nesterov momentum [38] was considered to improve MI-FGSM (N-MI-FGSM), which employs a "look ahead" gradient to more efficiently move away from suboptimal regions. Their method computes gradients at an estimated future position instead of the current position:

$$\tilde{x}_t^{adv} = x_t^{adv} + \alpha \cdot \mu \cdot g_t, \quad g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(\tilde{x}_t^{adv}, y)}{||\nabla_x J(\tilde{x}_t^{adv}, y)||_1}. \tag{6}$$

3.2 Critique of the Sign Function and Gradient Refinement

The aforementioned methods have established the fundamental framework for adversarial attacks. However, these approaches have been questioned by several researchers, particularly regarding the use of the sign function. Critics argue that this function discards the amplitude information present in the gradient, which may not consistently yield optimal adversarial examples.

Research into gradient function limitations appears in [39], where Taylor expansion analysis revealed directional inefficiencies of the sign operation, leading to their proposed Fast Gradient Non-sign Method (FGNM) with corresponding mathematical improvements. Complementary work by [40] established that direct gradient scaling techniques surpass sign-based approaches in transferability metrics while maintaining visual imperceptibility standards.

Researchers in [39] examined previous approaches, uncovering sign function inefficiencies and subsequently introducing FGNM. Their Taylor expansion analysis revealed limitations in sign-based methods while proposing mathematical improvements for attack optimization. Following this work, scholars developed direct gradient scaling techniques to replace sign functions [40]. Experiments verified these methods enhanced cross-model transferability of adversarial examples while maintaining necessary imperceptibility characteristics.

Raw Gradient Descent (RGD) [41] was proposed, which entirely removes the sign operation in attack framework. By reformulating the optimization problem from constrained to unconstrained, RGD utilizes the raw gradient directly:

$$x_{t+1}^{adv} = \Pi_{B_\epsilon(x)} \left(x_t^{adv} + \alpha \cdot \frac{\nabla_x J(x_t^{adv}, y)}{||\nabla_x J(x_t^{adv}, y)||_\infty} \right), \tag{7}$$

where $\Pi_{B_\epsilon(x)}$ is the projection operation onto the ϵ-ball centered at x. Their comprehensive experiments showed that RGD consistently outperforms PGD across various settings.

3.3 Advanced Optimization and Adaptive Dynamics

The integration of advanced optimization techniques into the well-established attack frameworks represents another advancement. Nadam optimizer [25] was incorporated into MI-FGSM, combining adaptive step size with look-ahead momentum to enhance effectiveness and transferability. The update rule is modified as:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign} \left(\frac{\beta_1 \cdot \hat{m}_t + (1 - \beta_1) \cdot g_t}{\sqrt{\hat{v}_t} + \epsilon} \right), \tag{8}$$

where m_t and v_t are the first and second moment estimates of the gradient, and β_1 and β_2 are the decay rates.

Researchers incorporated the AdaBelief optimizer into the attack framework, termed AB-FGSM [42]. This method dynamically adjusts α according to the estimated confidence in gradient calculations, resulting in notable cross-model transferability against both standard and adversarially-trained defensive systems. In a parallel development, the Adam Iterative Fast Gradient Method (AI-FGM) [43] harnesses Adam's momentum-based optimization and adaptive learning rates to construct more effective adversarial perturbations, demonstrating enhanced efficiency in penetrating model defenses.

To dismiss the adjustable step size, a general framework was presented [44], which can be implemented within aforementioned methods, such as MI-FGSM, N-MI-FGSM, etc. The proposed architecture can ensure better convergence and stability. The concept was extended through the introduction of a non-monotonic adaptive momentum coefficient combined with variable step-size methodology [45], accompanied by formal theoretical guarantees concerning regret limitations for convex functional spaces.

An innovative approach was introduced, employing variable step sizes that evolves throughout the attack process [46]. Their method specifically targets the common issue in iterative attacks such as I-FGSM, where transferability effectiveness tends to diminish with increasing iterations. By strategically adjusting and optimizing the utilization of gradient steps, their technique successfully preserves high levels of cross-model transferability across the entire optimization sequence, offering a solution to a significant limitation in traditional iterative attack methodologies.

"WITCHcraf" was proposed [47], a novel technique that incorporates randomized step sizes into the PGD framework. This approach effectively reduces initialization sensitivity while boosting performance efficiency without imposing additional computational burden. The research demonstrated that strategically randomizing the step sizes in the attack algorithm yields substantial improvements in successful attack rates. Despite its simplicity, this modification to traditional PGD proved to be remarkably effective in enhancing adversarial attack capabilities.

3.4 Transferability Enhancements and Spatial Considerations

Improving the transferability of adversarial examples across different models has been a central focus of recent research. Beyond the momentum-based methods discussed earlier, several innovative approaches have emerged.

The research work presented in [48] proposed a novel approach called Spatial Momentum Iterative FGSM (SMI-FGSM), which extends traditional adversarial attack methods by incorporating spatial-wise gradient accumulation alongside temporal momentum. This innovative technique takes into account contextual gradient information within images, resulting in more stable gradient update processes across various model architectures and datasets. By leveraging both spatial and temporal dimensions for momentum accumulation, SMI-FGSM demonstrates enhanced transferability capabilities compared to previous methods that relied solely on temporal momentum strategies.

The technique known as Scheduled Step Size and Dual Example (SD) was introduced in [49], which employs dynamic step size adjustment alongside dual examples to concentrate perturbations in proximity to benign samples. This methodology significantly enhances cross-model transferability by effectively preventing optimization processes from deviating excessively from the original sample distribution. Through this strategic approach, the researchers demonstrated improved attack efficiency while maintaining perturbation relevance.

The work presented in [50] introduces HE-MI-FGSM, a novel attack leveraging histogram equalization techniques to mitigate overfitting issues and boost black-box transferability. Through effective perturbation distribution regularization, this approach maintains its attacking capabilities while significantly improving generalization to unknown defensive models.

Research presented in [51] introduced a novel integration of Lookahead FGSM with Self-CutMix techniques to tackle the transferability limitations commonly observed under adversarial training scenarios. This approach enhances attack performance by intelligently utilizing internal patches from the target image, thereby preserving critical visual information and substantially improving success rates in black-box attack environments.

A different methodology described in [52] introduced the Gradient Relevance Attack framework, which implements adaptive direction correction mechanisms during the iterative perturbation generation process. This technique effectively minimizes gradient fluctuations by incorporating neighborhood information to adjust gradient relevance, resulting in remarkably high success rates against sophisticated defensive systems in black-box scenarios.

The work in [53] established an innovative Average Gradient-Based Adversarial Attack methodology that constructs and maintains a dynamic collection of adversarial examples throughout the attack process. By averaging gradient information across multiple iterations, this approach successfully mitigates the noise and instability issues inherent in traditional gradient updates, consequently enhancing the transferability of generated adversarial examples.

3.5 Feature-Region and Specialized Techniques

Numerous investigations have concentrated on specialized approaches for improving attack efficiency or minimizing detectability. The F-MIFGSM technique was developed [54], which confines perturbations to specific feature areas through the application of convolutional and deconvolutional neural network layers. By focusing on the most important regions within images, this method enhances attack concealment while simultaneously preserving high success metrics.

A different approach called Fast Gradient Scaled Method (FGScaledM) was developed [55], which implements gradient scaling to reduce perceptibility without compromising attack effectiveness. This methodology demonstrated that precisely controlled gradient scaling mechanisms could generate adversarial examples with substantially reduced visual detectability.

Subsequently, researchers established the Scale-Invariant PGD (SI-PGD) methodology [56], employing angular characteristics rather than logits to maintain consistent attack performance despite logit rescaling operations. This particular technique proved especially powerful against defensive mechanisms that rely on input transformation strategies or model scaling techniques.

3.6 Theoretical Advances and Comprehensive Frameworks

In the field of adversarial attacks, substantial progress has been made toward establishing theoretical underpinnings and consolidated frameworks for gradient-oriented attack methodologies. Researchers have proposed interpretative models that explain how diverse attack techniques can be conceptualized as variations of gradient descent with unique adaptation protocols [44]. Such frameworks not only deliver convergence assurances but also elucidate the interconnections between seemingly disparate attack approaches.

Further theoretical advancements have been achieved in understanding the mathematical properties of adaptive momentum techniques in adversarial contexts. A rigorous analysis establishing theoretical boundaries for regret in these methods has provided formal validation for the efficacy of using variable momentum parameters during optimization processes in adversarial settings [45].

The research community has additionally produced evaluative research comparing various attack implementations. A methodical evaluation of Fast Gradient Sign Method variants has been conducted, examining their operational mechanisms, inherent limitations, and performance impacts on ImageNet-trained ResNet-50 architectures [57]. These comparative investigations enhance our comprehension of the compromises inherent in different attack strategies and inform the development of more robust defensive measures.

3.7 Summary

Overall, PSAs have evolved from simple gradient-based methods to sophisticated optimization techniques that leverage advanced momentum, adaptive step sizes, and specialized refinements. The critique of the sign function has led to more effective gradient utilization, while transferability enhancements have improved attack success in black-box scenarios. These developments collectively represent a significant advancement in the field of adversarial machine learning, challenging the robustness of deep learning models in real-world applications.

4 Physically Realizable Attacks

Physically realizable adversarial attacks constitute a notable progression in the field of adversarial machine learning, transforming theoretical security vulnerabilities into concrete real-world threats. As shown in Figure 3 in contrast to digital-domain perturbations, attacks implemented in physical environments must sustain their effectiveness across various real-world conditions including changes in lighting, shifts in viewpoint, and distortions from printing processes. This section examines the progression and refinement of these attack methodologies and their deployment across various computer vision applications.

4.1 Early Foundations

The emergence of physically implementable adversarial attacks was pioneered through the development of the Robust Physical Perturbations (RP^2) algorithm [12]. This seminal research illustrated how strategically designed adhesive patterns could induce misclassification of traffic signage when observed across varying perspectives, distances, and illumination scenarios. The RP^2 methodology incorporated an optimization framework that considered real-world physical variations, imperfections in printing processes, and optical distortions, resulting in attack efficacy exceeding 80% during practical field evaluations. This groundbreaking investigation confirmed the feasibility of adversarial manipulations extending beyond computational environments and emphasized potential vulnerabilities in critical safety systems such as self-driving vehicles.

In parallel developments, researchers introduced techniques for generating physically robust adversarial three-dimensional objects [58]. This approach utilized the Expectation Over Transformation (EOT) framework to create adversarial examples maintaining effectiveness despite various transformations including rotational changes, spatial

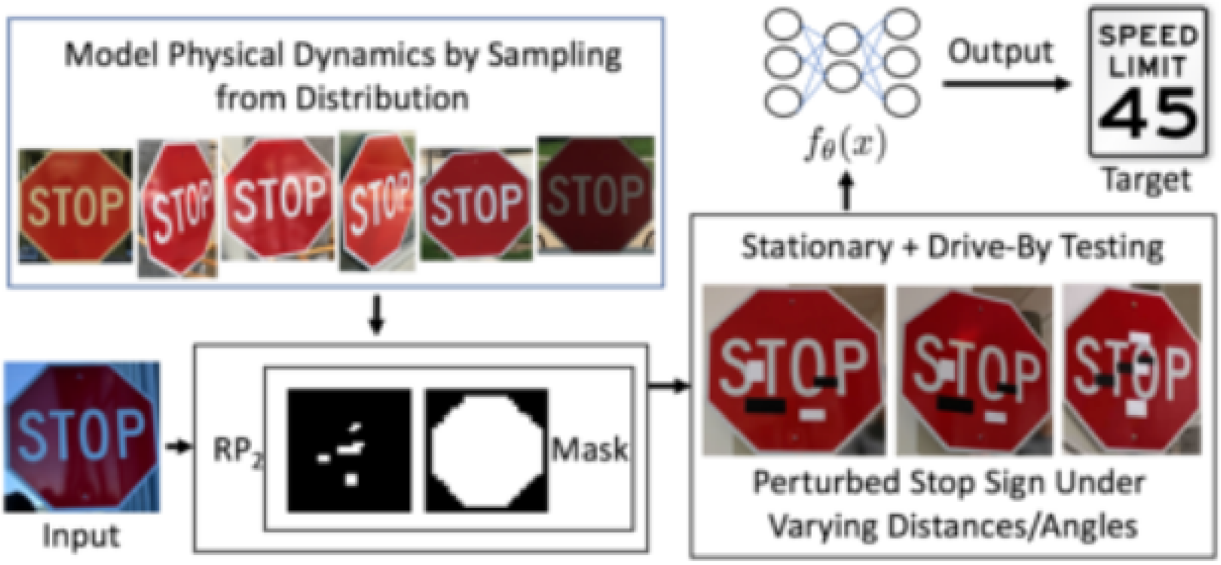

Model Physical Dynamics by Sampling from Distribution
Output
$f_\theta(x)$
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
Input
RP2
Mask
Perturbed Stop Sign Under Varying Distances/Angles

Figure 3: An example of physically realizable attacks [12], where the adversarial patch can easily fool the machine learning model.

repositioning, and illumination differences. By incorporating these transformation models during the optimization procedure, the research produced 3D-printed artifacts that consistently deceived classification systems regardless of observation angle. A notable example featured a turtle model consistently misidentified as a rifle from virtually any perspective, demonstrating how digital vulnerabilities successfully transfer to tangible physical objects.

4.2 Adversarial Patches and Stickers

Following these initial efforts, the domain of adversarial patches witnessed substantial advancement. Research presented in [59] introduced the concept of Adversarial Sticker attacks, which generated physically printable elements optimized for real-world applications. This methodology specifically tackled printing-related distortions and practical physical constraints, proving its efficacy across diverse applications including facial identification systems, traffic signage recognition, and image-based retrieval mechanisms. A particularly valuable contribution of this approach was its ability to maintain effectiveness in black-box scenarios, enabling successful attacks against systems without requiring access to their architectural specifications.

The field progressed significantly when researchers conducted comprehensive evaluations of the RP^2 approach across various environmental conditions [12]. Their investigations revealed how tactically positioned monochromatic adhesives could reliably cause autonomous systems to misinterpret stop signs as speed limit indicators, exposing critical vulnerabilities in vehicular automation technologies. This investigation highlighted the significance of environmental resilience, demonstrating that physical adversarial manipulations could sustain their effectiveness across varying observation distances, perspectives, and illumination settings.

A notable progression in adversarial methodologies emerged with the development of Universal Camouflage Patterns (UCP) [60]. Diverging from previous techniques that focused on specific targets, UCP generated adversarial patterns that remained effective against object detection systems across multiple contexts. This investigation introduced AttackScenes, a three-dimensional simulation environment that replicated real-world settings to validate attack performance, effectively connecting digital optimization processes with physical implementation challenges. The developed patterns achieved approximately 75% degradation in detection capabilities when applied to vehicles in authentic testing environments.

The quest for developing inconspicuous yet functional adversarial elements was addressed through the introduction of Naturalistic Physical Adversarial Patches (NPAP) [61]. This novel approach utilized generative adversarial networks to create visually convincing patches that appeared ordinary to human observers while maintaining their adversarial impact against object detection frameworks. By employing pre-established GANs such as BigGAN and StyleGAN, NPAP generated patches mimicking natural textures and designs, substantially enhancing the covertness of physical attacks while sustaining attack efficacy rates exceeding 80% against YOLOv3 detection systems.

In specialized application domains, research presented in [62] introduced adversarial patches specifically targeting monocular depth estimation technologies. These specially optimized elements were designed to disrupt depth perception in autonomous driving systems, inducing measurement errors of up to 6 meters with success rates of 93%. This work expanded the vulnerability landscape beyond traditional classification and detection tasks, demonstrating significant weaknesses in three-dimensional perception systems that are essential for navigation and obstacle detection functions.

4.3 3D Textures and Wearable Attacks

The transition from basic 2D adversarial patches to sophisticated 3D textured models marks a crucial advancement in physically realizable attack methodologies. Research introduced the Adversarial Textured 3D Meshes (AT3D) framework [63], which specifically targeted facial recognition technologies. This innovative approach tackled the intricate problem of designing adversarial textures that retain their effectiveness when applied to three-dimensional objects with diverse geometric properties. Through optimization of texture patterns that preserved their adversarial characteristics when mapped onto 3D facial meshes, researchers demonstrated highly effective attacks against commercial facial recognition frameworks under authentic operating environments, achieving ASR above 90% in real-world testing scenarios.

The development of wearable adversarial elements represents a particularly significant security concern due to their mobility and practical implementation potential. The Adversarial Clothing Textures (AdvCaT) methodology [64] addressed the specific challenges presented by non-rigid surfaces such as garments. This technique utilized advanced 3D modeling combined with Voronoi parameterization to generate robust textures capable of maintaining their adversarial properties despite the deformations experienced by clothing during normal body movements. To minimize the gap between digital simulations and physical implementation, this research incorporated fabrication constraints and color calibration protocols, ultimately achieving 90% attack efficacy against person detection systems across diverse environmental conditions.

The introduction of Legitimate Adversarial Patches (LAPs) [65] brought forward the critical concept of perceptual reasonableness alongside attack effectiveness. The research employed a dual-phase training methodology to create patches that appeared contextually natural while simultaneously retaining their capability to undermine object detection systems. By addressing the human perception aspect, this work overcame a significant limitation of previous adversarial patch designs that were visually conspicuous and readily identifiable as malicious elements. The LAPs approach demonstrated over 70% ASR while substantially reducing human suspicion metrics compared to conventional adversarial patch implementations.

Comprehensive evaluation research [66] provided thorough assessment of three-dimensional physical adversarial attacks, contrasting the performance of traditional 2D patches against complete 3D adversarial objects. This investigation examined critical variables including illumination variations, positional alterations, and material characteristics, establishing that 3D adversarial objects maintained superior ASR (85% compared to 67%) under varied observational conditions relative to their 2D counterparts. This systematic analysis established valuable benchmarks for assessing physical attack robustness and emphasized the enhanced adaptability of three-dimensional adversarial formations to environmental variations.

4.4 Dynamic and Optical Attacks

The field of adversarial attacks witnessed a significant advancement with the emergence of dynamic projection-based techniques. The concept of Short-lived Adversarial Perturbations (SLAP) [67] introduced a methodology utilizing projected light patterns to create temporary adversarial elements on objects. This innovative approach provided exceptional adaptability, enabling real-time adjustment to environmental fluctuations and moving targets. These projected adversarial elements achieved an 87% success rate in causing misclassification of traffic signage while leaving no enduring physical evidence, thus presenting substantial challenges for detection and countermeasures. The effectiveness of SLAP was demonstrated against both classification algorithms and object detection frameworks, with pattern optimization tailored for specific target models.

The optical attack landscape was further developed through EvilEye [68], an approach that implemented transparent display technology to generate dynamic optical disturbances. This research formalized the physical manifestation of digital attacks, establishing a versatile framework for real-time adversarial pattern deployment. Particularly noteworthy was EvilEye's performance in safety-critical applications, where it achieved ASR exceeding 90% against perception systems in surveillance and autonomous driving contexts. The method's capability to adjust perturbation intensity based on ambient conditions ensured consistent effectiveness across diverse lighting scenarios and distances, highlighting its remarkable environmental adaptability.

Another innovative direction in this domain emerged with Reflected Light Adversarial Attack (RFLA) [69], which employed strategically designed light reflection patterns to deceive computer vision systems. By optimizing reflection characteristics to maintain robustness under varied environmental conditions, RFLA demonstrated an impressive 99% ASR against image classification frameworks without requiring physical alterations to target objects. This methodology enabled nearly imperceptible attacks that presented exceptional challenges for detection and defense mechanisms, as the adversarial elements existed solely as ephemeral light patterns rather than permanent physical modifications.

The expansion into infrared-domain attacks broadened vulnerability exploitation to include thermal imaging technologies. Research introduced physically adversarial infrared patches [70] specifically designed to target thermal cameras deployed in surveillance systems and autonomous vehicles. These patches modified thermal distributions to generate adversarial patterns within the infrared spectrum, achieving over 80% success rates against detectors processing thermal imagery for pedestrian and vehicle identification. Through optimization of both patch configuration and positioning, this approach maintained effectiveness across temperature variations and viewing conditions, revealing vulnerabilities in systems engineered for operation in low-light environments and adverse weather scenarios.

4.5 Stealth Optimization Techniques

With the advancement of detection and defense strategies, the research community began to place greater emphasis on developing stealthier physical attack methodologies. The Dual Attention Suppression (DAS) attack was proposed [71], which employed a simultaneous optimization approach to circumvent both machine learning model attention mechanisms and human visual perception. This innovative technique generated contextually appropriate camouflage patterns that maintained adversarial effectiveness while appearing more natural to human observers, resulting in substantially improved imperceptibility metrics. Experimental validation demonstrated the DAS attack's efficacy across classification and object detection systems, with real-world physical implementations confirming its resilience under authentic operational conditions.

Further refinements in adversarial patch inconspicuousness emerged through sensitivity mapping techniques. A novel approach utilizing sensitivity maps was developed [72] to identify and exploit the most susceptible regions within object detection architectures while simultaneously reducing patch dimensions and visual prominence. By strategically concentrating adversarial perturbations at high-sensitivity model locations, this methodology achieved patch size reductions of approximately 60% while sustaining ASR exceeding 75%. This significant dimensional decrease substantially enhanced the stealthiness factor by rendering the adversarial elements considerably less detectable to human observers without sacrificing their effectiveness against computational systems.

Complementary research explored sparse adversarial patterning strategies. The Maximum Aggregated Region Sparseness (MARS) approach [73] was formulated to minimize and strategically localize attack regions on three-dimensional objects. Through the calculated placement of compact adversarial patterns at optimal positions, MARS achieved attack performance comparable to full-surface perturbations while substantially decreasing the modified object area. This methodology proved particularly advantageous for 3D object attacks, where precise positioning of small, unobtrusive adversarial elements at key viewpoints established an optimal balance between concealment and effectiveness.

Generative adversarial network (GAN) based natural-looking patches constituted another significant advancement in stealth optimization. As previously examined, the NPAP methodology [61] employed generative modeling techniques to create adversarial patches resembling authentic environmental textures. Subsequent investigations further refined this approach, developing implementations capable of remarkably convincing visual similarity to natural elements such as geological formations, plant matter, or textile patterns while preserving their adversarial characteristics. These naturalistic techniques substantially reduced human detection probability while maintaining high success rates against computer vision systems.

The introduction of distillation-enhanced optimization techniques for physical adversarial patches [80] represented an additional advancement in this domain. This methodology leveraged knowledge distillation principles to transfer adversarial properties between models, yielding patches with approximately 20% improved attack effectiveness alongside enhanced environmental integration capabilities. By optimizing patches to simultaneously emulate surrounding textures while maintaining their adversarial functionality, this approach achieved improved equilibrium between visual imperceptibility and attack efficacy, particularly for deployment scenarios involving visually complex environments.

4.6 Specialized Applications and Future Directions

When extended to specialized domains, physically realizable attacks have uncovered unique vulnerabilities specific to various contexts. Research in aerial detection systems has evaluated the effectiveness of adversarial patches against drone and satellite imaging technologies [81]. This investigation addressed the distinctive challenges presented by aerial viewpoints, including unusual observation angles and significant distances. The developed attack methods demonstrated resilience against variations in scale and environmental conditions inherent to aerial surveillance, illustrating how physical attacks can be adapted to specialized operational contexts.

Research has also introduced innovative attack vectors such as Out-of-Bounding-Box Triggers [82], which represent a covert approach that positions adversarial elements outside the conventional bounding box regions utilized by object detection algorithms. Through the implementation of feature guidance techniques and unified adversarial patch gradient descent methodology, researchers developed inconspicuous triggers capable of inducing detection failures while maintaining visual separation from targeted objects. This strategy achieved ASR exceeding 85% in real-world environments while substantially reducing human attention metrics compared to conventional adversarial patch implementations.

An emerging trend in this field involves the integration of multiple attack modalities. Contemporary approaches combine static physical patterns with dynamic light projections or incorporate adversarial textures with specialized materials affecting different sensing capabilities (including visual, infrared, and LiDAR systems). These multi-modal methodologies create significant obstacles for defensive mechanisms by simultaneously exploiting vulnerabilities across diverse sensing technologies and processing frameworks.

Current research increasingly emphasizes attack transferability across different computer vision tasks. While initial approaches targeted specific vision functions independently, recent investigations explore adversarial patterns that concurrently impact multiple vision operations. For instance, attacks initially designed for classification purposes have been successfully extended to compromise detection and segmentation functions without additional optimization, highlighting fundamental vulnerabilities shared across vision processing pipelines. This cross-task transferability substantially increases the potential threat of physical attacks in integrated vision systems.

Detailed literature reviews [83, 84, 85] have methodically classified these diverse attack strategies, identifying patterns, limitations, and future research directions in physically realizable adversarial attacks. These comprehensive analyses emphasize the accelerating technical advancements in attack methodologies, the broadening range of affected vision applications, and increasingly sophisticated concealment techniques. They also underscore persistent challenges, including effectiveness-concealment trade-offs, disparities between digital optimization and physical implementation, and requirements for standardized evaluation frameworks to assess system robustness against physical attacks.

In conclusion, physically realizable adversarial attacks have progressed from fundamental proof-of-concept demonstrations to sophisticated, concealed, and resilient attack vectors targeting various computer vision applications. From initial algorithms to advanced generative approaches, three-dimensional textured objects, and dynamic optical perturbations, these methodologies consistently demonstrate the vulnerability of vision systems to adversarial manipulation in real-world environments. The ongoing advancement of these techniques, combined with their increasing imperceptibility and robustness, presents substantial challenges for security and reliability in vision-based systems deployed in critical applications.

5 Latent-Space Attacks

Feature representation-based adversarial approaches form a sophisticated category of attack methods that target the internal feature spaces of deep neural networks rather than directly modifying input elements. As summarized in Figure 4, these techniques exploit vulnerabilities in high-level abstractions within models, enabling the generation of adversarial samples with enhanced cross-model transferability, greater semantic significance, and improved resilience against conventional defensive techniques.

5.1 Early Foundations

The investigation into LSAs gained momentum during 2017 – 2019, representing a paradigm shift from conventional pixel-space modifications toward manipulations within neural networks' internal representations. The groundbreaking work on LatentPoison [86] demonstrated the feasibility of generating adversarial examples through perturbations in deep variational autoencoders' latent representations. This approach implemented additive modifications directly to latent encodings, resulting in incorrect model predictions while preserving visual fidelity to the original inputs. This research established that exploiting latent vulnerabilities could facilitate more covert attacks with reduced perceptible artifacts.

Subsequent research expanded this direction when researchers proposed AdvGAN++ [87], which utilized generative adversarial networks to create adversarial examples through latent feature manipulation rather than direct image modification. This approach differentiated itself from earlier GAN-based techniques operating in pixel space by

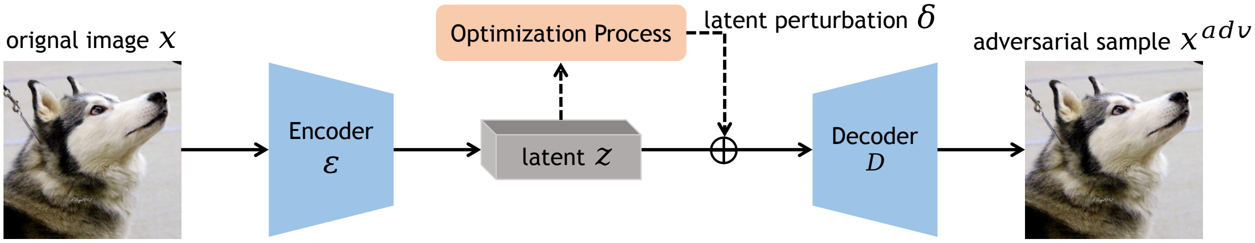


Figure 4: General paradigm of latent space attacks. The process flows from original samples through an encoder to latent space, where optimized perturbations are applied, followed by decoder transformation to generate deceptive yet visually similar adversarial examples. This attack methodology leverages the semantic structure of the latent space to achieve more efficient and targeted adversarial manipulations.

specifically targeting internal feature representations, thereby achieving superior attack effectiveness while maintaining perceptual similarity. This development illustrated the potential of generative frameworks to systematically explore and exploit vulnerabilities within latent spaces.

Concurrent developments included the introduction of Feature Space Perturbations [88], a methodology focused on improving adversarial transferability through strategic alignment of feature representations with target class characteristics. This technique identified the layers most susceptible to feature-level manipulations and demonstrated that latent-space perturbations could produce adversarial examples with enhanced cross-model generalization compared to conventional input-space approaches.

In response to these emerging threats, defensive strategies began to materialize, exemplified by the Latent Adversarial Defence (LAD) [89] framework, which targeted latent space vulnerabilities through decision boundary-oriented generation techniques. LAD operated by creating adversarial examples through feature perturbations along decision boundaries, subsequently reconstructing inputs for adversarial training purposes. This early defensive approach highlighted the growing recognition that securing latent spaces required specialized countermeasures beyond traditional input-space protection mechanisms.

5.2 Feature Distribution and Manipulation Approaches

The timeframe 2019 – 2021 witnessed remarkable advancements in LSA strategies, particularly those focusing on manipulating feature distributions and semantically significant representations. The Feature Distribution Attack (FDA) [90] emerged as a novel approach that deliberately altered class-specific feature distributions across multiple network layers, resulting in highly transferable adversarial examples. This methodology specifically targeted statistical characteristics within latent features, forcing them to deviate from their original class distribution toward designated target classes, thus demonstrating neural networks' susceptibility to distribution-based manipulations.

A significant breakthrough occurred with the introduction of LAFEAT [91], which pioneered gradient-oriented attacks specifically designed for latent feature manipulation. This technique utilized latent feature gradients to construct adversarial examples capable of circumventing even robust defense mechanisms, including models fortified through adversarial training. Research findings revealed that networks considered "robust" in input space remained vulnerable through their internal representations, highlighting a critical weakness in existing defensive approaches. By directly targeting intermediate features, this methodology achieved superior success rates against protected models compared to conventional attack vectors.

Research into semantic manipulation within latent spaces gained significant traction during this period. A framework for generating semantic adversarial examples through feature manipulation [92] utilized disentangled latent codes to create subtle yet interpretable adversarial examples. This approach identified and altered specific semantic attributes such as object shape, color, and textural properties within latent representations to induce misclassification while preserving visual coherence. Complementary research [93] presented methods for semantic adversarial perturbations using learned representations, employing generative adversarial networks to manipulate latent activations affecting high-level visual characteristics.

The vulnerability of internal layers in adversarially trained models received comprehensive examination through the development of LSA and Latent Adversarial Training (LAT) [94]. This research demonstrated that models specifically trained to withstand input-space perturbations nevertheless remained susceptible to manipulations of their latent representations. LAT addressed this vulnerability by implementing adversarial training directly within the latent space, establishing one of the earliest comprehensive defense mechanisms against LSAs.

Computational efficiency concerns in latent-space adversarial training were addressed through the Single-step Latent Adversarial Training (SLAT) [95] methodology. Unlike multi-step alternatives requiring substantial computational resources, SLAT manipulated latent gradients in a single step to enhance adversarial robustness efficiently, substantially reducing computational demands while maintaining defensive effectiveness. This innovation rendered latent-space defenses considerably more feasible for practical implementation in resource-constrained environments.

5.3 Generative Models and Semantic Control

The period from 2021 to 2023 witnessed significant advancements in LSAs, incorporating sophisticated generative models and enhanced semantic manipulation techniques. Research presented in [96] introduced a novel methodology for creating out-of-distribution adversarial examples through manipulation of latent space. This approach utilized β-VAEs to modify latent representations, producing adversarial samples that extended beyond conventional data distributions while preserving visual integrity. The work established how latent space modifications could produce semantically consistent examples that challenged traditional adversarial constraints.

Further developments emerged with the introduction of concept-based adversarial attacks as described in [97]. This methodology manipulated latent activations in deeper network layers to generate adversarial examples through semantic concept interpolation. By targeting high-level semantic features, this approach demonstrated effectiveness against both machine learning classifiers and human perception, illustrating how LSAs could simultaneously exploit vulnerabilities in both computational and human recognition systems.

A crucial breakthrough occurred with the application of diffusion models for latent space manipulations. The research in [98] established Semantic Adversarial Attacks using Diffusion Models, implementing Semantic Transformation (ST) and Latent Masking (LM) techniques for precise semantic control. The generative capabilities of diffusion models enabled fine-grained manipulation of semantic attributes while maintaining image fidelity. Compared to previous GAN-based methodologies, this diffusion-based framework demonstrated enhanced semantic richness and improved attack transferability.

The Latent Magic framework described in [99] investigated adversarial examples created within the semantic latent space of Stable Diffusion models. This research introduced innovative metrics for quantifying both attack effectiveness and cross-model transferability, demonstrating that perturbations in rich semantic latent spaces could generate adversarial examples with superior transferability compared to conventional pixel-space perturbations. Experimental evidence confirmed that latent representations from advanced generative models could be leveraged to create highly effective cross-architecture attacks.

Research on the GLASSE framework [100] integrated GANs with genetic algorithms to systematically explore latent spaces for adversarial example generation. This evolutionary approach navigated through latent manifolds to identify regions producing effective adversarial samples, achieving 82% success rates against external classification systems. The combination of generative modeling with evolutionary search techniques demonstrated how complex latent spaces could be methodically explored to locate vulnerable regions susceptible to adversarial manipulation.

Additional work in [101] enhanced semantic preservation in adversarial attacks by combining autoencoder architectures with genetic algorithms. This methodology extracted latent representations via autoencoders and subsequently applied genetic algorithms to modify these representations while maintaining semantic integrity. The approach preserved higher semantic fidelity compared to previous techniques while delivering competitive ASR, emphasizing the growing importance of ensuring adversarial perturbations remain semantically coherent with original inputs.

5.4 Advanced Geometric Understanding and Unrestricted Attacks

Recent advances in latent-space adversarial techniques have increasingly emphasized more nuanced geometric understanding of latent representations and unconstrained adversarial manifolds. Research by [102] introduced an innovative approach utilizing GAN-based modifications that thoroughly examines geometric characteristics of embedded features. This methodology strategically shifts feature representations toward incorrect class convex hulls, effectively circumventing input-space noise limitations while maintaining visual authenticity. Through comprehensive geometric evaluation, feature visualization techniques, and analysis of class activation patterns, the research illustrated how precisely targeted geometric manipulations within latent spaces yield highly successful adversarial outcomes.

The development of Direct Adversarial Latent Estimation (DALE) [103] offered new perspectives on decision boundary complexity assessment in black-box computational models. This framework employs variational autoencoders to generate and navigate through latent representations, producing adversarial samples for robustness evaluation. While primarily developed as an assessment tool rather than an attack mechanism, this investigation provided crucial insights regarding how decision boundaries in latent space influence model vulnerabilities, establishing theoretical foundations for understanding the variable effectiveness of latent perturbation techniques.

A breakthrough in unrestricted adversarial methodologies emerged with the introduction of Manifold-Aided Adversarial Examples [104]. This technique leverages supervised generative architectures to manipulate semantic characteristics within latent spaces through adversarial manifolds, generating examples that extend beyond conventional perturbation limitations. By effectively separating semantic from non-semantic attributes, the approach creates adversarial instances with legitimate natural semantics that remain difficult to identify through standard detection methods. This represents an evolution toward adversarial samples that maintain complete semantic integrity while achieving substantial attack effectiveness.

The introduction of Semantic-Consistent Attacks (SCA) [105] brought forth an exceptionally efficient framework for unrestricted adversarial attacks preserving semantic integrity. SCA implements diffusion inversion techniques to transform images into latent representations, modifies semantics under multimodal language model guidance, and ensures minimal distortion effects. This balanced approach combines adversarial optimization with semantic realism, demonstrating how contemporary generative models can be utilized to produce adversarial examples that maintain natural appearances while successfully deceiving target systems.

Research presented in [106] proposed an Adversarial Attack Algorithm utilizing Edge-Sketched Features from Latent Space (LSFAA), which successfully eliminates input-level iterative processes for more efficient adversarial generation. Through training specialized feature extraction mechanisms, LSFAA rapidly produces adversarial examples by manipulating edge-based latent characteristics, illustrating how LSA methodologies can be optimized for computational efficiency while sustaining high success rates.

5.5 Theoretical Insights and Properties

Moving beyond individual attack strategies, numerous research efforts have offered valuable theoretical perspectives on LSA characteristics. Research presented in [107] illustrated how adversaries targeting robust features could function as tools for interpretability, exposing model weaknesses through manipulation of high-level features in latent space. This investigation revealed that targeted, universal, and black-box latent space attacks could effectively highlight fragile semantic connections learned by neural networks, thus establishing a significant link between vulnerability to adversarial examples and model interpretability.

Several studies have thoroughly investigated the cross-model transferability of LSA. Research in [90] and [88] established that perturbations applied to latent representations generally demonstrate superior transferability across different model architectures compared to input-domain attacks. This enhanced transfer capability arises because diverse neural networks tend to develop comparable high-level feature representations despite architectural variations, making LSA approaches particularly potent in black-box attack scenarios.

The framework of Latent Manifold Adversarial Examples (LMAEs) introduced in [111] examined vulnerabilities in latent distributions from both local and global perspectives. By strategically modifying latent distributions and implementing manifold-aware adversarial training techniques, this approach significantly improved model resilience against diverse attack vectors. The theoretical analysis provided connections between latent-space vulnerabilities and the geometric characteristics of feature manifolds, offering explanations for why certain latent space regions exhibit greater susceptibility to adversarial manipulation.

Research documented in [112] showcased the generation of universal, physically-implementable adversarial features through latent representation manipulation. Utilizing deep generative models alongside specialized optimization procedures, this work crafted interpretable attacks exploiting feature-class associations within latent space, demonstrating how semantic-level modifications could translate into practical threats against computer vision systems.

The relationship between LSAs and semantic coherence has emerged as a consistent theme in contemporary research. The semantic adversarial attack methodology proposed in [113] employed parametric transformations in latent space, demonstrating how generative model latent spaces could be utilized to alter semantic attributes such as environmental conditions or facial characteristics. This research emphasized how semantically meaningful latent space transformations could generate adversarial outcomes that appear natural to human observers while consistently deceiving machine learning classifiers.

5.6 Future Directions

LSAs have evolved from simple perturbations in VAE and GAN latent codes to sophisticated manipulations leveraging advanced generative models and semantic control mechanisms. The progression from early works like LatentPoison [86] and AdvGAN++ [87] to recent approaches using diffusion models [98] and adversarial manifolds [104] reflects a deepening understanding of how internal feature representations can be exploited to create adversarial outcomes.

The field has seen several key trends emerge:

- increased focus on semantic coherence and naturalness in adversarial examples.
- exploitation of powerful generative models to navigate complex latent spaces.
- development of geometric understanding of latent vulnerabilities.
- enhanced transferability across model architectures.

These advances have established LSAs as a distinct and powerful class of adversarial techniques, capable of bypassing many defenses that focus exclusively on input-space perturbations.

Despite progress in defense mechanisms like Latent Adversarial Training [94] and boundary-guided approaches [89], latent-space vulnerabilities remain challenging to mitigate comprehensively. The continued evolution of this field suggests that understanding and addressing vulnerabilities in feature representations will remain a critical aspect of developing truly robust computer vision systems in the future.

6 Emerging Topic of Adversarial Attack in Computer Vision

6.1 Explore the Robustness Boundary of Authentication Systems

Authentication frameworks utilizing biological characteristics have undergone remarkable transformation throughout recent years, shifting from conventional pattern recognition techniques to advanced computational intelligence methodologies. While this progression has substantially improved identification precision and user experience, it simultaneously generates intricate security vulnerabilities that necessitate a fundamental reconsideration of protection strategies for biological identifiers. Central to these vulnerabilities exists what researchers term the non-replaceable attribute dilemma [114]. The unchangeable quality of physiological identifiers indicates that upon compromise, individuals cannot simply generate new ones as with alphanumeric credentials. Furthermore, these compromised physiological data points enable creation of effective oppositional instances, establishing significant risks to computational intelligence verification systems. This vulnerability has grown substantially more concerning within the computational learning paradigm.

6.1.1 The Irrevocability Paradox in Biometric Authentication

A critical security challenge with biometric authentication systems stems from their unchangeable nature, as documented by [114]. Security experts identify this as the “irrevocability paradox”: biometric identifiers including facial features, signature patterns, and fingerprint characteristics cannot be altered or replaced when compromised, unlike conventional authentication methods such as tokens or passwords. The unchanging property that provides convenience in biometric systems simultaneously constitutes their principal security weakness.

The concept of cancelable biometrics emerged as a solution to this vulnerability. This approach, introduced in foundational publications by [115, 116], involves applying systematic and reproducible alterations to biometric information through various mathematical operations. These included functional, polar, and Cartesian transformations applied to fingerprint data points. Instead of storing original biometric information, systems could retain the modified template, allowing for replacement with differently transformed versions if security became compromised. This methodology sought to add renewability while maintaining authentication performance.

The evolution of cancelable biometric techniques continued with the work of [117], who created revocable biotokens for fingerprints using binary field encoding strategies that simultaneously improved security aspects and recognition capabilities. Additional advancements came from [118] through the development of registration-independent fingerprint templates that prioritized template diversity and mathematical non-reversibility. A limitation of these approaches was their concentration on fixed security characteristics rather than resilience against developing attack methodologies.

6.1.2 Limitations of Static Revocation Mechanisms

The theoretical foundations of cancelable biometric systems have not translated effectively into practical applications. Research conducted by scholars [114] reveals substantial operational constraints involving unavoidable compromises among security, discriminability, and computational efficiency. A particularly significant flaw exists in their architectural design, which primarily addresses fixed-state threat models incapable of responding to the evolving nature of contemporary security challenges.

Such architectural limitations became increasingly evident as sophisticated adversarial methodologies developed. Academic literature [119, 120] offers a thorough analysis of biometric authentication frameworks within the context of adversarial machine learning principles. This research identifies critical vulnerabilities including authentication data corruption, synthetic identity presentation, and the substantial challenge of maintaining learning integrity within changeable threat landscapes. Their findings emphasize why protection mechanisms lacking adaptability invariably prove insufficient against strategically evolving security threats.

Further evidence supporting the insufficiency of non-adaptive cancellation protocols emerged through research [121] that introduced the concept of "biometric backdoors" as a specialized data poisoning technique targeting facial verification systems. By imperceptibly influencing the template modification protocols through strategically constructed adversarial inputs, researchers achieved sustained identity falsification capabilities that circumvented established protection measures. This vulnerability demonstrates how systems designed for continuous improvement through adaptation can unintentionally create novel security weaknesses when adversarial considerations are not fully integrated into their design architecture.

6.1.3 Machine Learning Era: Amplification of Vulnerabilities

The evolution toward machine learning technologies for biometric authentication, specifically systems utilizing deep learning structures, has considerably magnified challenges related to irrevocability. While providing enhanced recognition capabilities, these technologies have simultaneously generated novel security vulnerabilities which traditional protective frameworks cannot effectively counter.

Research published in [122] illustrates how generative adversarial networks successfully fabricate synthetic fingerprint patterns capable of deceiving deep learning recognition mechanisms. Complementary investigations documented in [123] demonstrate that adversarial perturbations, consisting of minimal alterations invisible to human examination, consistently compromise fingerprint authentication protocols. These discoveries underscore the substantial vulnerability of learning algorithms when confronted with sophisticated adversarial strategies.

Within handwritten signature authentication contexts, comprehensive analysis documented in [124] examined adversarial examples targeting offline signature verification frameworks. This research established that verification systems could be manipulated to incorrectly reject legitimate signatures or accept fabricated ones through deliberately constructed perturbations. Their analysis indicated that lower quality signature specimens demonstrated heightened susceptibility to adversarial manipulation, with error percentages escalating by approximately 49.19% under specific testing parameters. This weakness presents particular concerns for signature verification technologies since signature data inherently displays significant natural variation and inconsistency.

6.1.4 Dynamic Adversarial Environments and Adaptive Threats

Biometric security systems face significant challenges due to the continually transforming landscape of adversarial threats. Research [131] examining adaptive biometric frameworks reveals a crucial conflict between adaptation mechanisms designed to enhance recognition performance and those implemented for security purposes. When biometric systems are engineered to refresh templates for accommodating natural biometric variations, they may unintentionally create vulnerabilities that adaptive adversaries can systematically exploit.

This vulnerability has been substantiated through empirical research. A methodological approach for assessing risks associated with adversarial attacks on biometric technologies [132] illustrates how advanced attackers can leverage the relationships between various components within biometric architectures. Such exploitation creates sequential effects that conventional security assessments typically fail to identify. The investigation particularly emphasizes how template adaptation mechanisms remain susceptible to systematic adversarial manipulation across extended periods.

In response to increasingly complex security challenges, the scientific community has begun investigating more flexible defensive strategies. Recent contributions include a contextually aware framework for liveness verification in facial recognition systems [133] that incorporates situational data and surrounding conditions to dynamically adjust security configurations. This research constitutes significant progress toward developing biometric systems capable of adapting to emerging threat patterns, moving beyond traditional static protective measures.

6.1.5 Challenges of Low-Quality Samples in Signature Verification

Within the domain of biometric authentication systems, handwritten signatures constitute a particularly problematic category owing to their substantial variance and quality-dependent security implications. Contrary to physiological identifiers that maintain consistency, behavioral authentication factors such as handwritten signatures demonstrate natural fluctuations that simultaneously complicate recognition processes and security protocols.

Research conducted by investigators [134] illuminated these complexities, revealing that signature inconsistency functions as both advantage and liability: individuals can deliberately alter their signing patterns, providing natural replacement capabilities, while simultaneously creating openings for malicious exploitation of this inherent variability. The investigation highlighted how suboptimal signature acquisition substantially elevates false acceptance probabilities, thereby establishing security vulnerabilities that adversaries can leverage.

Subsequent quantitative analysis [124] further substantiated these concerns by revealing that antagonistic interventions against signature verification frameworks achieved notably greater success when utilizing or targeting signatures of inferior quality. Experimental evidence demonstrated that verification systems developed using pristine samples remained susceptible when processing degraded inputs, a situation frequently encountered in practical applications. Additionally, conventional protective strategies including adversarial augmentation displayed insufficient efficacy in addressing these vulnerabilities, underscoring requirements for more refined and responsive defensive methodologies.

6.1.6 Multimodal and Adaptive Defensive Strategies

The field has witnessed a significant shift toward multimodal biometric frameworks and dynamic protection mechanisms in addressing these security issues. Research contributions [135] have introduced an adaptive weighted graph methodology for creating multimodal cancelable biometric templates, illustrating how combining features at fundamental levels can simultaneously strengthen security and maintain template revocability. This methodology intelligently assigns weights to various biometric identifiers according to their quality assessments, thus offering resilience when sample quality fluctuates.

Contemporary studies [136] have performed extensive evaluations regarding how various integration techniques influence multimodal biometric system resilience when subjected to adversarial manipulation. Evidence suggests that integration at the input level typically offers greater protection compared to combining information at scoring or decision stages, although optimal integration strategies remain context-dependent. Significantly, this research emphasizes the necessity of incorporating adversarial resilience considerations during the initial system architecture phase rather than implementing protective measures retrospectively.

Concurrently, investigations [137] have explored the varying susceptibility levels across different biometric indicators when confronted with adversarial attacks. Research indicates certain identifiers (palmprints specifically) demonstrate heightened vulnerability compared to others (such as iris patterns). These observations suggest potential advantages in strategically selecting and differentially weighting biometric modalities to enhance overall system protection against adversarial threats.

6.1.7 The Expanding Robustness Boundary and Future Direction

Contemporary security systems continue to grapple with fundamental challenges despite advancements in countering adversarial threats and addressing irrevocability issues. The research community has yet to fully reconcile the static architecture of current revocation solutions with adversaries' evolutionary capabilities. Studies[132, 121] reveal how attackers continuously refine their methodologies, often circumventing the very protective mechanisms implemented to safeguard systems.

This fundamental incongruity becomes particularly evident when examining machine learning-powered signature authentication frameworks. These systems possess intricate model structures combined with inherent biometric variability, creating extensive vulnerability surfaces that remain insufficiently investigated. Conventional security assessment techniques predominantly examine fixed threat vectors, offering limited protection guarantees within this complex domain, thus necessitating more holistic and adaptable evaluation methodologies.

The scientific community must conduct thorough adversarial boundary testing to strengthen modern biometric frameworks. Through implementation of advanced attack protocols that evaluate systems under realistic and adaptive threat scenarios, researchers can uncover critical weaknesses and subsequently formulate more resilient protection mechanisms. This perspective conceptualizes security as a continuous refinement process rather than a fixed attribute—a consideration especially vital for signature verification technologies where authentication failures carry significant consequences.

As authentication systems incorporating machine learning continue their technological progression, we urgently require innovative frameworks capable of resolving the irrevocability paradox within evolving adversarial landscapes. Such advancement demands not only technical progress in cancelable biometrics and defensive countermeasures but also a paradigm shift in biometric security conceptualization—transitioning from static security assertions toward dynamic assurance frameworks capable of adapting at pace with emerging threats. Chapter 3 addresses these challenges by introducing our foundational framework for dynamic adversarial evaluation, providing adaptive testing methodologies that evolve alongside threat landscapes and establish the technical foundation for resilient biometric authentication systems.

6.2 Adversarial Protection against Malicious Generative Vision Models

The emergence and proliferation of sophisticated visual generative technologies have revolutionized our capacity to synthesize, modify, and create images. These technological innovations, while impressive in their capabilities, concurrently raise substantial concerns regarding security vulnerabilities and potential privacy violations. Such issues encompass unauthorized utilization of individual photographs, violations of copyright protections, and deliberate manipulation of visual media. This segment investigates the chronological progression of protective adversarial strategies designed to counteract unauthorized processing operations conducted by visual generative frameworks. Particular attention is directed toward models based on diffusion principles, given their current preeminence within the landscape of vision-oriented generative architectures.

6.2.1 Early Defenses Against Diffusion Models

During 2022 to 2023, with the increasing popularity of diffusion models, academic investigations began to explore adversarial methodologies for protecting visual content from unauthorized utilization. Among the initial contributions in this research area was the work presented in [138], which introduced AdvDM, a novel adversarial protection framework specifically engineered for diffusion architectures. This approach concentrated on interfering with the reverse diffusion sequence through strategic perturbations of latent representations, consequently preventing unauthorized extraction and replication of artistic elements. The significance of this research lies in its departure from conventional adversarial strategies, as it specifically targeted the sequential denoising operations fundamental to diffusion model functionality.

Subsequent research expanded this conceptual foundation with the introduction of MIST in [139], a sophisticated framework that substantially enhanced cross-model transferability of adversarial perturbations against diffusion systems. The authors of this study refined the adversarial objective function to ensure protective modifications remained robust across varied model structures and implementation parameters. Experimental validation confirmed that systematically constructed adversarial examples could effectively safeguard visual content from unauthorized synthesis operations while preserving perceptual quality.

In parallel developments, research documented in [140] introduced the Unlearnable Diffusion Perturbation (UDP) methodology, which represented a conceptual shift by focusing on preventing diffusion systems from effectively utilizing protected data during model training. This approach transitioned from inference-phase protection to training-phase safeguards, thereby addressing more comprehensive privacy and intellectual property considerations. The UDP technique implemented imperceptible adversarial modifications to visual content, rendering such materials unsuitable for training generative models such as Stable Diffusion, thus providing protection for creative styles and personal information against unauthorized appropriation.

6.2.2 Personalization and Identity Protection

The emergence of techniques for personalized text-to-image generation, exemplified by systems such as DreamBooth [141], has generated significant concerns regarding unauthorized utilization and identity theft. Research presented in [142] introduced Anti-DreamBooth, a protective framework that impairs personalized content generation through the application of adversarial perturbations to personal images. This system specifically targets the fundamental learning mechanisms that allow DreamBooth and comparable methods to establish connections between textual descriptions and visual identities, thus inhibiting unauthorized personalized content synthesis.

A different investigation documented in [143] established MetaCloak, a framework utilizing meta-learning principles to safeguard against unauthorized subject-based text-to-image generation. This approach generates resilient adversarial disturbances that maintain their effectiveness despite common image transformations including filtering operations and dimensional adjustments. The implementation of meta-learning enables optimization of protective perturbations across various generative models and textual prompt variations, constituting a notable improvement in defense mechanism durability.

The work presented in [144] introduced SimAC, which offers a straightforward approach to preventing personalization, specifically developed for facial privacy protection. This method incorporates time-step-sensitive adversarial noise coupled with feature-based optimization techniques to interrupt identity extraction processes in diffusion models. SimAC achieves an optimal balance between protective efficacy and retention of visual quality by simultaneously targeting frequency domains and encoder layers, rendering it particularly appropriate for protecting portrait imagery.

Addressing the need for instantaneous identity protection, the research in [145] established Real-time Identity Defense (RID), a system generating protective perturbations through a single-forward-pass neural network architecture. This efficiency-oriented solution facilitates practical implementation for extensive image protection scenarios, particularly on social media platforms requiring immediate processing capabilities. The RID system demonstrates that effective adversarial protection measures can be implemented without excessive computational requirements, thus addressing a major obstacle to widespread adoption of such protective technologies.

6.2.3 Robust and Transferable Protection

Research into enhanced adversarial defense mechanisms emerged following discoveries that initial defenses exhibited insufficient cross-model transferability and susceptibility to preprocessing manipulations. The introduction of Prompt-Agnostic Perturbations (PAP) by researchers [146] represented a substantial advancement in this domain. This approach utilized Laplace Approximation to conceptualize prompt distributions, generating robust perturbations that maintained efficacy across diverse contexts and prompt configurations. The PAP methodology substantially enhanced the cross-model applicability of adversarial protections, effectively addressing a fundamental weakness inherent in previous techniques.

Protection through Score Distillation Sampling was investigated in scholarly work [147], with particular focus on targeting latent diffusion models' encoder components to establish effective adversarial safeguards. This strategy achieved computational efficiency while sustaining resistance against unauthorized utilization. Concurrent research [148] established an Adversarial Decoupling Augmentation Framework (ADAF) that implemented text-associated augmentations, creating consistent protections against various malicious input prompts, particularly beneficial for facial privacy preservation applications.

Addressing the critical issue of maintaining defensive integrity against preprocessing techniques and adversarial purification methods, investigators [149] formulated DiffVax, an optimization-independent framework designed to shield images from unauthorized diffusion-based modifications. This innovation facilitated instantaneous, scalable image protection that retained effectiveness despite preprocessing operations such as compression and dimensional alterations, constituting a noteworthy advancement in practical adversarial defense implementation.

6.2.4 Hybrid Approaches and Watermarking

In response to the constraints of purely adversarial techniques, the research community shifted toward integrated solutions that merge adversarial modifications with additional safeguarding mechanisms. Research presented in [150] established a framework for Watermark-embedded Adversarial Examples, utilizing conditional generative adversarial networks to create adversarial samples that compel diffusion models to generate outputs containing visible watermarks. This integrated protection framework enables simultaneous copyright verification and quality reduction, tackling multiple dimensions of unauthorized utilization concurrently.

In a parallel research direction, scholars in [151] formulated Robust Invisible Watermarking (RIW), a method leveraging adversarial principles to ensure watermark persistence throughout transformations based on diffusion processes. Their methodology maintained extraction accuracy rates of 96% following content modification, substantiating the effectiveness of permanent identity integration as a supplementary protective strategy.

6.2.5 Evaluation and Limitations

The assessment of protective measures' efficacy and constraints has evolved into a systematic research focus as this domain advanced. A framework known as IMPRESS was established by researchers [152] to assess how resilient imperceptible perturbations are against unauthorized utilization of data in generative AI based on diffusion models. Their investigation uncovered that current methodologies remain susceptible to purification approaches, which accentuates the persistent difficulty in developing genuinely robust protective systems.

Further investigations have raised significant concerns regarding the capability of protective perturbations to shield personal information from exploitation. Research conducted on Stable Diffusion [153] introduced the purification methodology GrIDPure, which demonstrates the capacity to circumvent adversarial protections while maintaining image utility, thus exposing fundamental weaknesses in contemporary defensive techniques. Comparable findings were reported in additional studies [154], which illustrated how techniques focusing on feature alignment, such as INSIGHT, could effectively neutralize protective perturbations that are imperceptible, thereby enabling diffusion models to regenerate features from images that were supposedly protected. These research outcomes emphasize the continuous evolution within the defense-attack ecosystem and highlight the requirement for protection mechanisms that adapt progressively.

6.2.6 Recent Advancements and Specialized Defenses

The literature reveals an evolution toward context-specific protective mechanisms tailored to unique security threats and application environments. Research presented in [155] established DiffusionGuard as a defensive framework that counters unauthorized diffusion-based image manipulation by focusing on initial diffusion phases while implementing mask-augmentation techniques. This innovation specifically tackles the prevention of illicit partial content alterations, addressing a significant security vulnerability in real-world implementations.

A methodology termed Dual-Domain Anti-Personalization (DDAP) was formulated in [156], introducing concurrent disturbances across both spatial and frequency realms to impede texture and detail acquisition in customized text-to-image generation systems. The simultaneous application across multiple domains enables DDAP to attain enhanced resilience in comparison with conventional single-domain techniques, thus validating the efficacy of comprehensive protection frameworks.

Concerning the targeted safeguarding of visual elements, the VCPro framework outlined in [157] implements strategic adversarial modifications to combat counterfeit image creation and stylistic reproduction while maintaining fundamental perceptual integrity. This nuanced protection strategy facilitates adaptable implementation contexts wherein security requirements focus exclusively on particular components within visual content.

6.2.7 Critical Gaps and Future Directions

Although considerable advancements have occurred in developing protective measures against adversarial attacks on diffusion models, numerous important research challenges remain unexplored. A significant oversight pertains to neural style transfer applications, which pose unique threats to artists' intellectual property rights. While substantial research addresses generalized image manipulation and personalized content generation, the scholarly community has insufficiently investigated protective techniques specifically tailored for style transfer scenarios. This gap is directly addressed in Chapter 4, where we develop targeted protective techniques specifically for neural style transfer scenarios, providing a comprehensive protection framework for artistic intellectual property.

An additional limitation involves computational efficiency and knowledge requirements. Contemporary adversarial techniques targeting diffusion frameworks typically necessitate comprehensive understanding of model internals, particularly regarding UNet architectural specifications. This requirement substantially restricts cross-model applicability and imposes excessive computational burdens, rendering practical implementation challenging. The field would benefit significantly from methodologies maintaining effectiveness while requiring minimal model information. These limitations are tackled through our grey-box framework in Chapter 5, which requires minimal model knowledge while maintaining effectiveness.

These unaddressed challenges represent fertile ground for subsequent scholarly investigation, potentially yielding more comprehensive and implementable safeguards against unauthorized utilization of visual assets by generative AI systems.

7 Conclusion

7.1 Summary of Key Findings

This comprehensive literature review has examined the evolving landscape of adversarial attacks in computer vision systems, revealing significant developments across multiple attack paradigms and application domains. Our analysis demonstrates that adversarial vulnerabilities represent both fundamental challenges to the reliability of deep learning systems and valuable tools for security assessment and protection.

Methodological Evolution: The field has progressed from simple gradient-based perturbations (FGSM, PGD) to sophisticated optimization techniques incorporating momentum, adaptive step sizes, and advanced transferability mechanisms. Pixel-space attacks have evolved beyond basic sign-function approaches to embrace raw gradient methods and specialized optimization frameworks, significantly improving both effectiveness and cross-model transferability.

Physical World Implications: Physically realizable attacks have successfully bridged the gap between digital vulnerabilities and real-world threats. The development from basic adversarial patches to sophisticated 3D textures, wearable attacks, and dynamic optical perturbations demonstrates the practical severity of adversarial vulnerabilities in deployed systems.

Latent-Space Sophistication: Latent-space attacks have emerged as a particularly powerful paradigm, leveraging the semantic structure of internal representations to create more transferable and semantically meaningful adversarial examples. The progression from simple VAE manipulations to sophisticated diffusion-model-based attacks reflects the field's adaptation to advancing generative technologies.

Dual Nature of Applications: Perhaps most significantly, our review reveals the dual nature of adversarial techniques. While these methods expose critical vulnerabilities in computer vision systems, they also serve as valuable tools for robustness assessment and protective applications, including biometric security evaluation and prevention of unauthorized content generation.

7.2 Challenges and Future Directions

Building upon the significant contributions outlined in the previous sections, several critical challenges and potential directions for future research in adversarial machine learning for computer vision remain.

7.2.1 Enhancing Transferability in Adversarial Attacks

One of the persistent challenges in adversarial machine learning is the transferability of attacks across different model architectures and domains. While recent work has demonstrated promising results in specific contexts, such as the high transferability of attacks across various Latent Diffusion Models, several issues remain unresolved:

- **Cross-architecture transferability**: The effectiveness of adversarial attacks often diminishes when applied to models with fundamentally different architectures [158]. Future research should focus on identifying and exploiting common vulnerabilities across diverse neural network designs, potentially through more abstract representations of model behavior rather than architecture-specific features.
- **Domain adaptation for adversarial attacks**: Developing techniques that can adapt adversarial perturbations to new domains without requiring extensive modification would significantly enhance the practical utility of these methods. This could involve learning domain-invariant features [159] that are consistently susceptible to perturbation across different contexts.
- **Model-agnostic protection mechanisms**: As demonstrated by the Posterior Collapse Attack, targeting fundamental components common across multiple model architectures (such as Variational Autoencoders (VAE) in LDMs) offers a promising direction for creating more universal protection mechanisms. Extending this approach to other shared components or principles in machine learning models represents an important avenue for future work.

7.2.2 Balancing Perturbation Imperceptibility and Attack Effectiveness

Throughout our research, a consistent challenge has been striking the optimal balance between the imperceptibility of adversarial perturbations and their effectiveness in achieving the desired outcome. This trade-off varies across different applications:

- **Domain-adaptive perturbation constraints**: Future research should explore methods for dynamically adjusting perturbation constraints based on image content, viewing conditions, and application context. For instance, images with high-frequency textures may tolerate larger perturbations without perceptible changes compared to smoother images.
- **Perceptually guided adversarial optimization**: Incorporating more sophisticated models of human visual perception into adversarial optimization processes could lead to perturbations that more effectively exploit the gaps between machine and human vision. This may include leveraging insights from psychophysics to identify perturbations that are imperceptible to humans while maximally disrupting machine learning models.

7.2.3 Extending to Emerging AI Paradigms

As AI continues to evolve, new paradigms and architectures emerge, presenting both challenges and opportunities for adversarial machine learning:

- **Foundation models and large vision models**: The rise of foundation models [160, 161] and large vision models [162] introduces new vulnerabilities and protection requirements. Exploring adversarial attacks and defenses in these contexts, particularly in multi-modal systems that combine vision and language, represents an important direction for future research.

- **Adversarial robustness in self-supervised learning**: As self-supervised learning becomes more prevalent in computer vision, understanding the unique vulnerabilities and robustness properties of these approaches compared to supervised learning is crucial for developing effective protection mechanisms.
- **Quantum machine learning**: Looking further ahead, the emergence of quantum machine learning algorithms [163] may fundamentally change the landscape of adversarial attacks and defenses. Investigating the implications of quantum computing for adversarial robustness represents a long-term research direction.

7.3 Implications and Future Outlook

The findings of this review have several important implications for the computer vision community. First, the rapid evolution of attack methodologies demonstrates that adversarial robustness cannot be treated as a static property but requires continuous adaptation to emerging threats. Second, the success of physically realizable attacks underscores the critical need for robustness evaluation under real-world conditions rather than purely digital settings.

The emerging applications of adversarial techniques for protective purposes represent a paradigm shift in how we conceptualize these methods. Rather than viewing adversarial attacks solely as security threats, the community is increasingly recognizing their potential as tools for privacy protection and intellectual property safeguarding.

Future research should prioritize the development of more generalized attack and defense frameworks that can adapt to rapidly evolving AI technologies. This includes investigating the fundamental principles underlying adversarial vulnerabilities and developing theoretical frameworks that can guide both attack development and defense strategies.

The field would benefit from increased collaboration between adversarial ML researchers and domain experts in critical applications such as healthcare, autonomous systems, and security. Such interdisciplinary approaches could lead to more practical and effective solutions that address real-world deployment challenges.

As computer vision systems become increasingly integrated into society's critical infrastructure, ensuring their adversarial robustness is not merely an academic pursuit but a societal imperative. The continued advancement of both attack and defense methodologies will be essential for building trustworthy AI systems that can operate reliably in adversarial environments.

References

[1] Yanhao Jia, Xinyi Wu, Hao Li, Qinglin Zhang, Yuxiao Hu, Shuai Zhao, and Wenqi Fan. Uni-retrieval: A multi-style retrieval framework for stem's education. *arXiv preprint arXiv:2502.05863*, 2025.

[2] Yanhao Jia, Xinyi Wu, Qinglin Zhang, Yiran Qin, Luwei Xiao, and Shuai Zhao. Towards robust evaluation of stem education: Leveraging mllms in project-based learning. *arXiv preprint arXiv:2505.17050*, 2025.

[3] Zhongliang Guo, Ognjen Arandjelović, David Reid, Yaxiong Lei, and Jochen Büttner. A siamese transformer network for zero-shot ancient coin classification. *Journal of Imaging*, 9(6):107, 2023.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[6] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[7] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[8] Zhongliang Guo, Chun Tong Lei, Lei Fang, Shuai Zhao, Yifei Qian, Jingyu Lin, Zeyu Wang, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. A grey-box attack against latent diffusion model-based image editing by posterior collapse. *arXiv preprint arXiv:2408.10901*, 2024.

[9] Zhongliang Guo. *Building trustworthy computer vision: adversarial techniques for robustness assessment and misuse prevention*. PhD thesis, The University of St Andrews, 2025.

[10] Zhongliang Guo, Junhao Dong, Yifei Qian, Kaixuan Wang, Weiye Li, Ziheng Guo, Yuheng Wang, Yanli Li, Ognjen Arandjelović, and Lei Fang. Artwork protection against neural style transfer using locally adaptive adversarial color attack. In *ECAI 2024*, pages 1414–1421. IOS Press, 2024.

[11] Zhongliang Guo, Yifei Qian, Shuai Zhao, Junhao Dong, Yanli Li, Ognjen Arandjelović, Lei Fang, and Chun Pong Lau. Artwork protection against unauthorized neural style transfer and aesthetic color distance metric. *Pattern Recognition*, page 112105, 2025.

[12] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.

[14] Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. *arXiv preprint arXiv:2101.08030*, 2021.

[15] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593, 2019.

[16] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[17] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[18] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pages 1615–1631, 2018.

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

[20] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. In *Advances in Neural Information Processing Systems*, volume 35, pages 7546–7560, 2022.

[21] Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. Instruct2Attack: Language-guided semantic adversarial attacks. *arXiv preprint arXiv:2311.15551*, 2023.

[22] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178, 2020.

[23] Shilin Qiu, Qihe Liu, Shijie Zhou, and Chunjiang Wu. Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, 9(5):909, 2019.

[24] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Qikun Zhang, Yuzhi Zhang, Yanling Shao, Mengqi Liu, Jianyong Li, Junling Yuan, and Ruifang Wang. Boosting adversarial attacks with nadam optimizer. *Electronics*, 12(6):1464, 2023.

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[27] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[28] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 2023.

[29] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020.

[30] Jiawei Zhang, Linyi Li, Huichen Li, Xiaolu Zhang, Shuang Yang, and Bo Li. Progressive-scale boundary blackbox attack via projective gradient estimation. In *International Conference on Machine Learning*, pages 12479–12490. PMLR, 2021.

[31] Kunyu Wang, Juluan Shi, and Wenxuan Wang. LFAA: Crafting transferable targeted adversarial examples with low-frequency perturbations. In *ECAI 2023*, pages 2483–2490. IOS Press, 2023.

[32] Qi Liang, Qiang Li, and Song Yang. LP-GAN: Learning perturbations based on generative adversarial networks for point cloud adversarial attacks. *Image and Vision Computing*, 120:104370, 2022.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[36] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018.

[38] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

[39] Yaya Cheng, Jingkuan Song, Xiaosu Zhu, Qilong Zhang, Lianli Gao, and Heng Tao Shen. Fast gradient non-sign methods. *arXiv preprint arXiv:2110.12734*, 2021.

[40] Zheng Yuan, Jie Zhang, Zhaoyan Jiang, Liangliang Li, and Shiguang Shan. Adaptive perturbation for adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[41] Junjie Yang, Tianlong Chen, Xuxi Chen, Zhangyang Wang, and Yingbin Liang. Rethinking PGD attack: Is sign function necessary? *CoRR*, 2023.

[42] Yixiang Wang, Jiqiang Liu, and Xiaolin Chang. Generalizing adversarial examples by adabelief optimizer. *arXiv preprint arXiv:2101.09930*, 2021.

[43] Ronghui Zhou. Study of optimiser-based enhancement of adversarial attacks on neural networks. In *2024 International Conference on Interactive Intelligent Systems and Techniques (IIST)*, pages 740–747, 2024.

[44] Wei Tao, Lei Bao, Sheng Long, Gaowei Wu, and Qing Tao. Adapting step-size: A unified perspective to analyze and improve gradient-based methods for adversarial attacks. *arXiv preprint arXiv:2301.11546*, 2023.

[45] Sheng Long, Wei Tao, LI Shuohao, Jun Lei, and Jun Zhang. On the convergence of an adaptive momentum method for adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14132–14140, 2024.

[46] Qidong Huang, Leiji Lu, Jun Chen, and Lei Bao. Using dynamically changing step sizes to increase the success rate of adversarial attacks. In *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pages 1071–1076. IEEE, 2024.

[47] Ping-Yeh Chiang, Jonas Geiping, Micah Goldblum, Tom Goldstein, Renkun Ni, Steven Reich, and Ali Shafahi. WITCHcraft: Efficient pgd attacks with random step size. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3747–3751. IEEE, 2020.

[48] Guoqiu Wang, Huanqian Yan, and Xingxing Wei. Enhancing transferability of adversarial examples with spatial momentum. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 593–604. Springer, 2022.

[49] Zeliang Zhang, Peihan Liu, Xiaosen Wang, and Chenliang Xu. Improving adversarial transferability with scheduled step size and dual example. *arXiv preprint arXiv:2301.12968*, 2023.

[50] Youqing Fang, Jingwen Jia, Yuhai Yang, and Wan-Li Lyu. Adversarial example generation method based on probability histogram equalization. In *2023 42nd Chinese Control Conference (CCC)*, pages 7951–7958, 2023.

[51] Donggon Jang, Sanghyeok Son, and Dae-Shik Kim. Strengthening the transferability of adversarial examples using advanced looking ahead and Self-CutMix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 147–154, 2022.

[52] Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4741–4750, 2023.

[53] Chen Wan, Fangjun Huang, and Xianfeng Zhao. Average gradient-based adversarial attack. *IEEE Transactions on Multimedia*, 25:9572–9585, 2023.

[54] Sijie Liu, Zhixiang Zhang, Xian Zhang, and Haojun Feng. F-MIFGSM: adversarial attack algorithm for the feature region. In *2020 IEEE 9th joint international information technology and artificial intelligence conference (ITAIC)*, volume 9, pages 2164–2170. IEEE, 2020.

[55] Zhefeng Xu, Zhijian Luo, and Jinlong Mu. Fast gradient scaled method for generating adversarial examples. In *Proceedings of the 2022 6th International Conference on Innovation in Artificial Intelligence*, pages 189–193, 2022.

[56] Mengting Xu, Tao Zhang, Zhongnian Li, and Daoqiang Zhang. Scale-invariant adversarial attack for evaluating and enhancing adversarial defenses. *arXiv preprint arXiv:2201.12527*, 2022.

[57] Aarti Lad, Ruchi Bhale, and Shlok Belgamwar. Fast gradient sign method (FGSM) variants in white box settings: A comparative study. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 382–386. IEEE, 2024.

[58] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, pages 284–293. PMLR, 2018.

[59] Xingxing Wei, Ying Guo, and Jie Yu. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2711–2725, 2022.

[60] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[61] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7848–7857, 2021.

[62] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *European Conference on Computer Vision*, pages 514–532. Springer, 2022.

[63] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. Towards effective adversarial textured 3D meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4128, 2023.

[64] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16975–16984, 2023.

[65] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5307–5315, 2021.

[66] Mohammad Zarei, Chris Ward, Josh Harguess, and Marshal Aiken. Adversarial barrel! an evaluation of 3D physical adversarial attacks. In *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6, 2022.

[67] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1865–1882, 2021.

[68] Yi Han, Matthew Chan, Eric Wengrowski, Zhuohuan Li, Nils Ole Tippenhauer, Mani Srivastava, Saman Zonouz, and Luis Garcia. Why don't you clean your glasses? perception attacks with dynamic optical perturbations. *arXiv preprint arXiv:2307.13131*, 2023.

[69] Donghua Wang, Wen Yao, Tingsong Jiang, Chao Li, and Xiaoqian Chen. RFLA: A stealthy reflected light adversarial attack in the physical world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4455–4465, 2023.

[70] Xingxing Wei, Jie Yu, and Yao Huang. Physically adversarial infrared patches with learnable shapes and locations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12334–12342, 2023.

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

[72] Ce Wang and Qianmu Li. Stealthy adversarial patch for evading object detectors based on sensitivity maps. In *2024 IEEE Cyber Science and Technology Congress (CyberSciTech)*, pages 322–328. IEEE, 2024.

[73] Ling Zhao, Xun Lv, Lili Zhu, Binyan Luo, Hang Cao, Jiahao Cui, Haifeng Li, and Jian Peng. A local adversarial attack with a maximum aggregated region sparseness strategy for 3d objects. *Journal of Imaging*, 11(1):25, 2025.

[74] Yifei Qian, Xiaopeng Hong, Zhongliang Guo, Ognjen Arandjelović, and Carl R Donovan. Semi-supervised crowd counting with contextual modeling: Facilitating holistic understanding of crowd scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[75] Yanli Li, Jifei Hu, Zhongliang Guo, Nan Yang, Huaming Chen, Dong Yuan, and Weiping Ding. Threats and defenses in federated learning life cycle: A comprehensive survey and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[76] Yifei Qian, Liangfei Zhang, Zhongliang Guo, Xiaopeng Hong, Ognjen Arandjelović, and Carl R Donovan. Perspective-assisted prototype-based learning for semi-supervised crowd counting. *Pattern Recognition*, 158:111073, 2025.

[77] Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Feng Yichao, Fengjun Pan, and Anh Tuan Luu. A survey of recent backdoor attacks and defenses in large language models. *Transactions on Machine Learning Research*, 2025. Survey Certification.

[78] Man Hu, Yatao Yang, Deng Pan, Zhongliang Guo, Luwei Xiao, Deyu Lin, and Shuai Zhao. Syntactic paraphrase-based synthetic data generation for backdoor attacks against chinese language models. *Information Fusion*, page 103376, 2025.

[79] Yuqi Li, Yanli Li, Kai Zhang, Fuyuan Zhang, Chuanguang Yang, Zhongliang Guo, Weiping Ding, and Tingwen Huang. Achieving fair medical image segmentation in foundation models with adversarial visual prompt tuning. *Information Sciences*, page 122501, 2025.

[80] Wei Liu, Yonglin Wu, Chaoqun Li, Zhuodong Liu, and Huanqian Yan. Distillation-enhanced physical adversarial attacks. *arXiv preprint arXiv:2501.02232*, 2025.

[81] Jiawei Lian, Shaohui Mei, Shun Zhang, and Mingyang Ma. Benchmarking adversarial patch against aerial detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[82] Tao Lin, Lijia Yu, Gaojie Jin, Renjue Li, Peng Wu, and Lijun Zhang. Out-of-bounding-box triggers: A stealthy approach to cheat object detectors. In *European Conference on Computer Vision*, pages 269–287. Springer, 2024.

[83] Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9797–9817, 2024.

[84] Xingxing Wei, Bangzheng Pu, Jiefan Lu, and Baoyuan Wu. Visually adversarial attacks and defenses in the physical world: A survey. *arXiv preprint arXiv:2211.01671*, 2022.

[85] Amira Guesmi, Muhammad Abdullah Hanif, Bassem Ouni, and Muhammad Shafique. Physical adversarial attacks for camera-based smart systems: Current trends, categorization, applications, research challenges, and future outlook. *IEEE Access*, 11:109617–109668, 2023.

[86] Antonia Creswell, Anil A Bharath, and Biswa Sengupta. Latentpoison-adversarial attacks on the latent space. *arXiv preprint arXiv:1711.02879*, 2017.

[87] Surgan Jandial, Puneet Mangla, Sakshi Varshney, and Vineeth Balasubramanian. AdvGAN++: Harnessing latent layers for adversary generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2045–2048, 2019.

[88] Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[89] Xiaowei Zhou, Ivor W Tsang, and Jie Yin. Latent adversarial defence with boundary-guided generation. *arXiv preprint arXiv:1907.07001*, 2019.

[90] Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *International Conference on Learning Representations*, 2020.

[91] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. LAFEAT: Piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5735–5745, 2021.

[92] Shuo Wang, Shangyu Chen, Tianle Chen, Surya Nepal, Carsten Rudolph, and Marthie Grobler. Generating semantic adversarial examples via feature manipulation in latent space. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[93] Isaac Dunn, Tom Melham, and Daniel Kroening. Semantic adversarial perturbations using learnt representations. *arXiv preprint arXiv:2001.11055*, 2020.

[94] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Harnessing the vulnerability of latent layers in adversarially trained models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2779–2785, 2019.

[95] Geon Yeong Park and Sang Wan Lee. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7758–7767, 2021.

[96] Ujjwal Upadhyay and Prerana Mukherjee. Generating out of distribution adversarial attack using latent space poisoning. *IEEE Signal Processing Letters*, 28:523–527, 2021.

[97] Johannes Schneider and Giovanni Apruzzese. Concept-based adversarial attacks: Tricking humans and classifiers alike. In *2022 IEEE Security and Privacy Workshops (SPW)*, pages 66–72. IEEE, 2022.

[98] Chenan Wang, Jinhao Duan, Chaowei Xiao, Edward Kim, Matthew c Stamm, and Kaidi Xu. Semantic adversarial attacks via diffusion models. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[99] BoYang Zheng. Latent magic: An investigation into adversarial examples crafted in the semantic latent space. *arXiv preprint arXiv:2305.12906*, 2023.

[100] Luana Clare and João Correia. Generating adversarial examples through latent space exploration of generative adversarial networks. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 1760–1767, 2023.

[101] Xinyi Wang, Simon Yusuf Enoch, and Dan Dongseong Kim. Semantic preserving adversarial attack generation with autoencoder and genetic algorithm. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 80–85, 2022.

[102] Nitish Shukla and Sudipta Banerjee. Generating adversarial attacks in the latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 730–739, 2023.

[103] Ashley S Dale and Lauren Christopher. Direct adversarial latent estimation to evaluate decision boundary complexity in black box models. *IEEE Transactions on Artificial Intelligence*, 2024.

[104] Shuai Li, Xiaoyu Jiang, and Xiaoguang Ma. Transcending adversarial perturbations: Manifold-aided adversarial examples with legitimate semantics. *arXiv preprint arXiv:2402.03095*, 2024.

[105] Zihao Pan, Weibin Wu, Yuhang Cao, and Zibin Zheng. SCA: Highly efficient semantic-consistent unrestricted adversarial attack. *CoRR*, 2024.

[106] Yangjie Cao, Chenxi Zhu, Haobo Wang, and Yan Zhuang. An adversarial attack algorithm based on edge-sketched feature from latent space. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 723–728. IEEE, 2022.

[107] Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33093–33106. Curran Associates, Inc., 2022.

[108] Ziheng Guo, Zhongliang Guo, Oggie Arandelovic, and Andrea di Falco. Generative model for multiple-purpose inverse design and forward prediction of disordered waveguides in linear and nonlinear regimes. In *Machine Learning in Photonics*, volume 13017, page 1301702. SPIE, 2024.

[109] Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P. Pound. T2icount: Enhancing cross-modal understanding for zero-shot counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 25336–25345, 2025.

[110] Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 24331–24340, 2025.

[111] Zhuang Qian, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Improving model robustness with latent distribution locally and globally. *arXiv preprint arXiv:2107.04401*, 2021.

[112] Stephen Casper, Max Nadeau, and Gabriel Kreiman. One thing to fool them all: Generating interpretable, universal, and physically-realizable adversarial features. *Preprint*, 2021.

[113] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[114] Anil K. Jain, Karthik Nandakumar, and Abhishek Nagar. Biometric template security. *EURASIP J. Adv. Signal Process.*, 2008, 2008.

[115] Nalini Ratha, Jonathan Connell, Ruud M Bolle, and Sharat Chikkerur. Cancelable biometrics: A case study in fingerprints. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 370–373. IEEE, 2006.

[116] Nalini K Ratha, Sharat Chikkerur, Jonathan H Connell, and Ruud M Bolle. Generating cancelable fingerprint templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):561–572, 2007.

[117] Terrance E Boult, Walter J Scheirer, and Robert Woodworth. Revocable fingerprint biotokens: Accuracy and security analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[118] Zhe Jin, Andrew Beng Jin Teoh, Thian Song Ong, and Connie Tee. A revocable fingerprint template for security and privacy preserving. *KSII Transactions on Internet and Information Systems (TIIS)*, 4(6):1327–1342, 2010.

[119] Battista Biggio, Paolo Russu, Luca Didaci, Fabio Roli, et al. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, 32(5):31–41, 2015.

[120] Zhongliang Guo, Weiye Li, Yifei Qian, Ognjen Arandjelovic, and Lei Fang. A white-box false positive adversarial attack method on contrastive loss based offline handwritten signature verification models. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909. PMLR, 2024.

[121] Giulio Lovisotto, Simon Eberz, and Ivan Martinovic. Biometric backdoors: A poisoning attack against unsupervised template updating. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 184–197. IEEE, 2020.

[122] Hee won Kwon, Jea-Won Nam, Joongheon Kim, and Youn Kyu Lee. Generative adversarial attacks on fingerprint recognition systems. In *2021 International Conference on Information Networking (ICOIN)*, pages 483–485. IEEE, 2021.

[123] Stefano Marrone and Carlo Sansone. Adversarial perturbations against fingerprint based authentication systems. In *2019 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2019.

[124] Luiz G Hafemann, Robert Sabourin, and Luiz S Oliveira. Characterizing and evaluating adversarial examples for offline handwritten signature verification. *IEEE Transactions on Information Forensics and Security*, 14(8):2153–2166, 2019.

[125] Xinxin Liu, Zhongliang Guo, Siyuan Huang, and Chun Pong Lau. Mmad-purify: A precision-optimized framework for efficient and scalable multi-modal attacks. *arXiv preprint arXiv:2410.14089*, 2024.

[126] Shuai Zhao, Leilei Gan, Zhongliang Guo, Xiaobao Wu, Luwei Xiao, Xiaoyu Xu, Cong-Duy Nguyen, and Luu Anh Tuan. Weak-to-strong backdoor attack for large language models. *arXiv preprint arXiv:2409.17946*, 2024.

[127] Yuqi Li, Xingyou Lin, Kai Zhang, Chuanguang Yang, Zhongliang Guo, Jianping Gou, and Yanli Li. Fedkd-hybrid: Federated hybrid knowledge distillation for lithography hotspot detection. *arXiv preprint arXiv:2501.04066*, 2025.

[128] Chun Tong Lei, Zhongliang Guo, Hon Chung Lee, Minh Quoc Duong, and Chun Pong Lau. Towards more transferable adversarial attack in black-box manner. *arXiv preprint arXiv:2505.18097*, 2025.

[129] Hon Ming Yam, Zhongliang Guo, and Chun Pong Lau. My face is mine, not yours: Facial protection against diffusion model face swapping. *arXiv preprint arXiv:2505.15336*, 2025.

[130] Shuai Zhao, Yulin Zhang, Luwei Xiao, Xinyi Wu, Yanhao Jia, Zhongliang Guo, Xiaobao Wu, Cong-Duy Nguyen, Guoming Zhang, and Anh Tuan Luu. Affective-roptester: Capability and bias analysis of llms in predicting retinopathy of prematurity. *arXiv preprint arXiv:2507.05816*, 2025.

[131] Norman Poh, Ajita Rattani, and Fabio Roli. Critical analysis of adaptive biometric systems. *IET Biometrics*, 1(4):179–187, 2012.

[132] Seong Hee Park, Soo-Hyun Lee, Min Young Lim, Pyo Min Hong, and Youn Kyu Lee. A comprehensive risk analysis method for adversarial attacks on biometric authentication systems. *IEEE Access*, 2024.

[133] Emma Lavens, Davy Preuveneers, and Wouter Joosen. Mitigating undesired interactions between liveness detection components in biometric authentication. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–8, 2023.

[134] Tasmina Islam. *Increasing Reliability and Security in Handwritten Signature Biometrics*. University of Kent (United Kingdom), 2017.

[135] Gurjit Singh Walia, Gaurav Jain, Nipun Bansal, and Kuldeep Singh. Adaptive weighted graph approach to generate multimodal cancelable biometric templates. *IEEE Transactions on Information Forensics and Security*, 15:1945–1958, 2019.

[136] Shaima M Alghamdi, Salma Kammoun Jarraya, and Faris Kateb. Enhancing security in multimodal biometric fusion: Analyzing adversarial attacks. *IEEE Access*, 2024.

[137] MyeongHoe Lee, JunHo Yoon, and Chang Choi. Adversarial attack vulnerability for multi-biometric authentication system. *Expert Systems*, 41(10):e13655, 2024.

[138] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pages 20763–20786. PMLR, 2023.

[139] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

[140] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

[141] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[142] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2116–2127, 2023.

[143] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. MetaCloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24219–24228, 2024.

[144] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12047–12056, 2024.

[145] Hanzhong Guo, Shen Nie, Chao Du, Tianyu Pang, Hao Sun, and Chongxuan Li. Real-time identity defenses against malicious personalization of diffusion models. *arXiv preprint arXiv:2412.09844*, 2024.

[146] Cong Wan, Yuhang He, Xiang Song, and Yihong Gong. Prompt-agnostic adversarial perturbation for customized diffusion models. *Advances in Neural Information Processing Systems*, 37:136576–136619, 2024.

[147] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *International Conference on Learning Representations*, 2024.

[148] Ruijia Wu, Yuhang Wang, Huafeng Shi, Zhipeng Yu, Yichao Wu, and Ding Liang. Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. *arXiv preprint arXiv:2305.03980*, 2023.

[149] Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P Chinchali, and James M Rehg. Optimization-free image immunization against diffusion-based editing. *arXiv preprint arXiv:2411.17957*, 2024.

[150] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded adversarial examples for copyright protection against diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24420–24430, 2024.

[151] Mingtian Tan, Tianhao Wang, and Somesh Jha. A somewhat robust image watermark against diffusion-based editing models. *arXiv preprint arXiv:2311.13713*, 2023.

[152] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023.

[153] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24398–24407, 2024.

[154] Shengwei An, Lu Yan, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, and Xiangyu Zhang. Rethinking the invisible protection against unauthorized image usage in stable diffusion. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3621–3638, 2024.

[155] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, 2025.

[156] Jing Yang, Runping Xi, Yingxin Lai, Xun Lin, and Zitong Yu. DDAP: Dual-domain anti-personalization against text-to-image diffusion models. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024.

[157] Xiaoyue Mi, Fan Tang, Juan Cao, Peng Li, and Yang Liu. Visual-friendly concept protection via selective adversarial perturbations. *CoRR*, 2024.

[158] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[159] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research*, 2022.

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

[161] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[162] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22861–22872, 2024.

[163] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.