# A Review of Deep Face Adversarial Sample Generation Techniques

*Chenxi Cui*

School of Robotics Engineering, Harbin Far East Institute of Technology, Harbin, China

**Abstract.** With the wide application of deep learning in face recognition systems, it has achieved remarkable results in the fields of identity verification and security monitoring. However, it has been found that such systems are extremely vulnerable to adversarial samples, and the attacker only needs to add tiny, imperceptible perturbations to make the model output misclassified. The paper systematically reviews the definition, classification and typical generation methods of face adversarial samples, covering both physical and digital domain attacks. It further discusses the challenges of black-box and white-box attacks on detection systems and the effectiveness of mainstream defense means such as adversarial training and residual denoising networks. Finally, the deficiencies of current detection techniques in terms of generalization capability, deployment efficiency and physical attack defense are analyzed, and the future development direction of building a unified evaluation system and a multimodal robust detection framework is proposed. This research provides theoretical support and practical paths for the optimization of security protection and adversarial sample detection strategies for face recognition systems.

## 1 Introduction

Deep learning advancements have established face recognition as a key computer vision/AI technology, widely used in identity authentication, public security, finance, and social media. Modern systems leverage deep models (notably CNNs) for feature extraction, achieving superhuman accuracy on benchmarks. However, studies show critical security vulnerabilities in these models, particularly susceptibility to adversarial examples, raising concerns about real-world robustness.

Adversarial examples refer to input images that have been subtly modified by adding imperceptible perturbations, with the specific intent of misleading deep learning models into making incorrect predictions. Although these altered images appear nearly identical to the original ones to the human eye, they can cause significant deviations in model output. For instance, an attacker may wear accessories such as glasses or hats embedded with adversarial patterns to deceive face recognition systems, achieving identity obfuscation or

*yunse@ldy.edu.rs*

evading detection without raising human suspicion. This phenomenon not only exposes the fragility of conventional face recognition algorithms but also presents a severe security risk in critical domains such as surveillance, financial risk management, and judicial oversight.

In recent years, research on adversarial face samples have made notable progress. Sharif et al. first proposed the "adversarial glasses" attack, which optimizes eyeglass frame patterns to cause face recognition systems to misidentify the wearer as another individual [1]. This method is effective not only in digital environments but also in physical settings through printed patterns and real-world usage. Komkov et al. extended this idea with the "adversarial hat," applying perturbed patterns to the hat brim. Through affine transformations and multi-angle simulations, the method significantly lowers identity similarity scores, demonstrating strong transferability and concealment in real scenarios [2].

In digital attacks, Goodfellow et al. introduced the Fast Gradient Sign Method (FGSM), which leverages model gradients to quickly generate perturbations, revealing the linear vulnerabilities of deep networks [3]. Carlini and Wagner advanced this with the C&W attack, which crafts imperceptible adversarial samples with minimal distortion, bypassing multiple defenses [4]. Chen et al. further proposed the EAD method, incorporating elastic-net regularization into the C&W framework to improve perturbation sparsity, stealth, and attack success rate [5].

Although significant progress has been made in adversarial sample generation, attack strategies, and defense mechanisms, several critical challenges remain unresolved. First, the feasibility and stability of physical-world attacks lack a systematic theoretical foundation, and there is no unified evaluation framework to assess the effectiveness of perturbations under varying conditions such as angles, lighting, and backgrounds. Second, most detection methods rely heavily on specific datasets for training, resulting in poor generalization and limited robustness against novel or composite attacks. Third, mainstream models still face performance bottlenecks in terms of lightweight deployment and real-time responsiveness, hindering their applicability in edge computing and mobile scenarios.

Building on this background, this paper presents a systematic review of adversarial samples in face recognition, focusing on the following key aspects: First, it defines and categorizes adversarial face samples, and examines attack targets and methodologies for both face recognition and face detection systems. Second, it provides a detailed overview of mainstream adversarial sample generation techniques across both physical and digital domains, comparing their implementation strategies, applicable scenarios, and attack effectiveness. Third, it analyzes the security threats posed by black-box and white-box attack models, and reviews representative research along with current limitations in detection and defense mechanisms. By systematically synthesizing existing work on adversarial face samples, this study aims to offer researchers a comprehensive technical roadmap and a holistic understanding of the field, thereby contributing to the development of more robust, secure, and interpretable face recognition systems.

## 2 Overview of Face Confrontation Samples

### 2.1Definition of Face Confrontation Samples

Face adversarial samples are facial images intentionally modified with carefully crafted perturbations that mislead deep learning models into producing incorrect classifications, while remaining visually recognizable to human observers. Mathematically, given the original image $x$ and classification of objectives $y$，Confrontation Sample $x'$ Needs to be satisfied，As shown in Equation 1:

$$x' = x + \delta, \text{where } \| \delta \|_p \leq \epsilon \quad (1)$$

Among them,δis the added perturbation，$\epsilon$Controlling the size of the perturbation，$p$ is usually taken as 2 (Euclidean paradigm) or $\infty$ (maximum number of paradigms).

## 2.2 Primary Target: Disruption of Artificial Intelligence Systems vs. Disruption of Human Sensory Systems

Face adversarial samples can be broadly categorized into two main types. The first involves interference with artificial intelligence systems, where adversarial perturbations are designed to mislead automated face recognition systems into misclassification or failure to recognize a target. The second pertains to interference with human perception, where techniques such as deepfake generation are employed to create highly realistic fake faces, making it difficult for human observers to distinguish between real and falsified information. Figure 1 presents a hierarchical framework for the classification of face adversarial samples, highlighting these two primary branches “Interference with AI systems” and “Interference with human perception systems” which correspond to the principal targets of adversarial attacks.

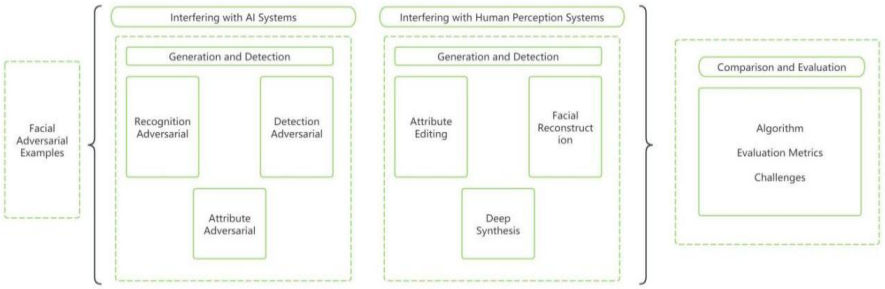


Fig. 1 Classification of face confrontation samples (Picture credit: Original)

## 2.3 Classification of Face Adversarial Samples

### *2.3.1 Face Recognition Adversarial Sample*

Face recognition systems authenticate identities by extracting and matching facial features based on biometric data. Adversarial attacks typically take two forms: adding subtle, imperceptible perturbations or embedding adversarial patches that preserve visual integrity. These attacks cause misclassification or reduced confidence by biasing model outputs, exposing weaknesses in robustness. While posing security risks, they also provide insights for improving model generalization through reverse optimization.

### *2.3.2 Face Detection Against Sample*

Face detection is a core module in computer vision, used to localize facial features and regress bounding boxes via feature pyramid networks and region proposal algorithms. Adversarial attacks mainly target deep learning-based face detectors and general object

detection models through gradient-based perturbations or adversarial patches. These attacks cause shifted boxes, missing detections, or sharp drops in confidence (often below the 60% threshold), revealing weaknesses in model robustness. Such vulnerabilities pose security risks in applications like surveillance and liveness detection, while also offering data for adversarial training and defense research.

# 3 Generation Techniques for Face Confrontation Samples

Adversarial attacks on face recognition systems have rapidly advanced, extending from digital to physical domains. Generation methods can be classified into physical attacks, global digital attacks, and hybrid approaches, all designed to fool recognition systems while remaining imperceptible to humans.

## 3.1 Face Recognition System Attack

### *3.1.1 Physically Realized Adversarial Face Samples*

Researchers have found that printing special adversarial patterns on eyeglass frames can effectively trick neural networks into identifying the attacker as another identity. Sharif et al. showed that wearing eyeglasses with specific color patterns can successfully bypass commercial face recognition systems, allowing the attacker to be incorrectly classified as the specified target identity [1]. An attack method against facial recognition systems is presented, which centers on designing and printing special eyeglass frames to spoof the system. The attacks are categorized into two types; Dodging and Impersonation. Where in Impersonation, the attacker wishes to find perturbation $\mathrm{r}$， make the image $\mathrm{x}+\mathrm{r}$ is categorized as a target class $\mathrm{ct}$, As shown in Equation 2:

$$\mathrm{argrmin}(\mathrm{softmaxloss}(\mathrm{f}(\mathrm{x}+\mathrm{r}),\mathrm{ct})+\kappa\cdot\mathrm{TV}(\mathrm{r})+\kappa'\cdot\mathrm{NPS}(\mathrm{r})) \quad (2)$$

where softmaxloss is the classification loss function，$\mathrm{TV}(\mathrm{r})$ is the total variance of the perturbation used for the smoothness constraints，$\mathrm{NPS}(\mathrm{r})$is the non-printability score used to ensure that perturbations are printable，$\kappa$ and $\kappa'$ is the equilibrium parameter.

And for Dodging, the attacker wants to find the perturbation $\mathrm{r}$，make the image $\mathrm{x}+\mathrm{r}$ Not classified as primitive $\mathrm{cx}$, As shown in Equation 3:

$$\mathrm{argminr}(-\mathrm{softmaxloss}(\mathrm{f}(\mathrm{x}+\mathrm{r}),\mathrm{cx})+\kappa\cdot\mathrm{TV}(\mathrm{r})+\kappa'\cdot\mathrm{NPS}(\mathrm{r})) \quad (3)$$

Experiments confirm that specially designed eyeglass frames can effectively spoof face recognition systems in both digital and physical settings. Figure 2 visually illustrates the attack process and results, highlighting the practicality of physical attacks and the need for robust defenses.

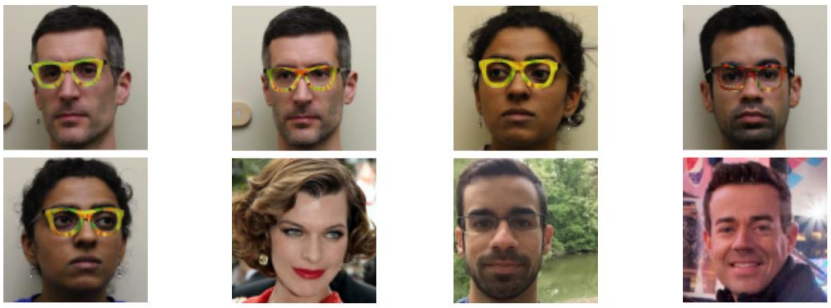

Fig. 2 Schematic diagram of face confrontation samples [1]

Komkov et al. developed an attack on facial recognition systems through an adversarial hat (AdvHat) that reduces similarity scores between captured facial images and authentic reference images, thereby preventing authentication success [2]. AdvHat's objective is to generate a rectangular printable pattern that, when affixed to a hat, causes facial recognition systems to calculate similarity scores below predetermined authentication thresholds. Specifically, attackers aim to produce a wearable sticker that, when captured alongside the face, renders the facial recognition system unable to match the image to legitimate reference templates. The methodology accounts for realistic physical deformation by modeling the sticker's bending and rotation when worn on a hat, implemented through 3D affine and parabolic transformations. The 3D affine transformation is expressed as Equation 4

$$z = a \cdot x^2 \tag{4}$$

The spatial transformation layer is then used to project the sticker to a high resolution face image (600×600). Randomly adjust parameters such as position and scaling during the projection process to simulate small changes in the actual shot. The STL is again used to perform the affine transformation (alignment, cropping). Random perturbations such as translation, rotation are introduced in the transformation to enhance the generalization of adversarial perturbations. Finally adversarial stickers are generated by gradient descent to minimize the real face similarity and keep the perturbation smooth. Experimental results demonstrate AdvHat reduces cosine similarity scores by over 0.5 (versus typical authentication thresholds of 0.3-0.8) under controlled conditions. The attack remains effective despite head orientation and lighting variations, and transfers successfully to other architectures including LResNet50E-IR. Figure 3 illustrates the stickers' disruptive impact through comparative experiments and quantitative analysis, providing critical empirical evidence for developing defensive countermeasures.

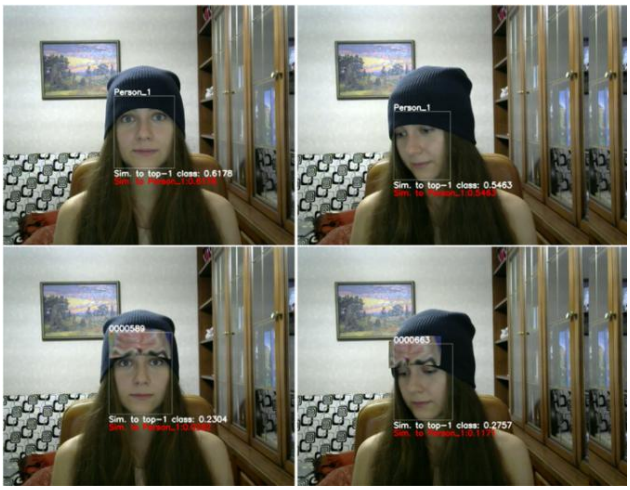


Fig. 3 Stickers affixed to hats reduce the similarity to ground truth classes [2]

Face Painting and Stickers: Applying high contrast colors or random noise to specific areas of the face can disrupt the model's extraction of key features such as the eyes, nose, and mouth, thereby reducing recognition accuracy. Physical stickers, such as adversarial stickers can be placed on areas such as the forehead or cheeks to trick the system into failing to correctly detect the location of the face. Eykholt et al. found that commercial face recognition systems can be effectively tricked by wearing stickers with random noise [6]. The literature describes an algorithm called RP2 for generating adversarial perturbations that can mislead deep learning visual classification models in the physical world. The algorithm generates adversarial perturbations through an optimization method and takes into account environmental variations in the physical world to ensure the robustness of the perturbations in real-world applications, the, The optimization problem can be expressed as Equation 5:

$$m_{\delta}(\lambda \parallel \delta \parallel + J(f(x+\delta), y^{*})) \tag{5}$$

A mask Mx, matching the input image size, is used to restrict perturbations to specific regions of the target object. It contains ones in regions to be perturbed and zeros elsewhere. L1 and L2 regularization are applied to guide the mask toward vulnerable areas. Two classifiers, LISA-CNN and GTSRB-CNN, are built following a standard crop-adjust-classify pipeline and trained on the LISA and GTSRB datasets, respectively.

### *3.1.2 Global Digital Adversarial Face Sample*

The Fast Gradient Symbol Method (FGSM) proposed by Goodfellow et al. is representative of the work on gradient optimization attacks, revealing the vulnerability of neural networks to confrontation due to their linear nature [3]. The theoretical basis of the method lies in the fact that the generation of adversarial samples originates from the linear perturbation sensitivity of the model inputs, and this assumption is verified by cross-model generalization experiments. FGSM generates perturbations by calculating the gradient symbols of the loss function, which achieves efficient adversarial sample generation in a single-step optimization. In addition, the method incorporates adversarial samples into the

training process for the first time, which not only improves the robustness of the model, but also improves the generalization performance through the regularization effect. The modular process of "gradient computation → perturbation → construction → sample generation" of the FGSM has had a profound impact on the subsequent research, and has become a benchmark tool for evaluating the safety of the model. Figure 4 below presents the key steps of FGSM in a modular way to visualize the core link between perturbation generation and adversarial sample output.

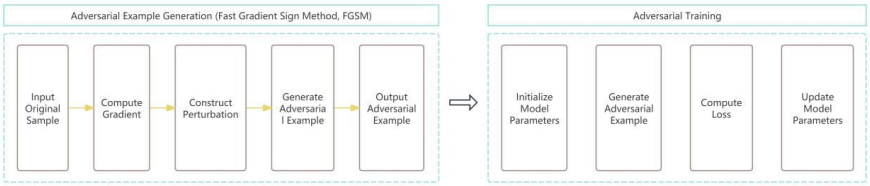


Fig. 4 Technical flowchart of FGSM (Picture credit: Original)

Iterative Gradient Attack (I-FGSM): Kurakin et al. proposed I-FGSM to obtain stronger attack by multi-step optimization [7]. Kurakin et al. found that even after complex transformations of physical media (printer noise, camera distortion, lighting variations, etc.), a significant percentage of adversarial samples were still misclassified [7]. For example, antagonistic samples generated using the fast method in the ε=16， About 66% of the samples still lead to Top-1 classification errors. Inside, we introduce the generation of adversarial samples through multiple iterations of FGSM (Fast Gradient Symbol Method), where the input is updated with small step sizes at each iteration and the maximum magnitude of the perturbation is limited by a cropping operation (L∞ paradigm constraint). From the original image $x = x_0$ Start, the update formula for each step is: Starting from, as shown in Equation 6:

$$\mathrm{Xn} + 1 = \mathrm{ClipX}, \epsilon\{\mathrm{Xn} + \alpha \cdot \mathrm{sign}(\nabla \mathrm{XJ}(\mathrm{Xn}, \mathrm{ytrue}) \quad (6)$$

The number of iterations is $\min(\epsilon + 4, 1.25\epsilon)$， Dynamically adjusted to balance the efficiency and attack strength. Test different $\epsilon$ values (e.g., 2, 4, 8, 16) to verify the effectiveness of the adversarial samples. The survival rate of the antagonistic samples in the physical world is tested by printing, photographing and image transformation. Finally compared with other methods, the principle of Iterative Gradient Symbol Method (I-FGSM), experimental validation and comparison results with other attack methods are illustrated as shown in Table 1 below:

**Table 1** Comparison with other methods

| methodologies | perturbation property | physical robustness | Applicable Scenarios |
|---|---|---|---|
| FGSM | Single-step large disturbances with significant noise | High | Fast generation, high noise tolerance scenarios |
| I-FGSM | Multi-step fine perturbation for concealment | Relatively low | Physical attacks that require high stealth |

| | | | |
|---|---|---|---|
| I-FGSM | Directed attack, perturbation targeted | Relatively low low | Specific Target Misdirection Scenario |

## 3.2 Face Detection Against Attacks

### *3.2.1: Physical Domain Attacks*

Physical adversarial attacks are meticulously engineered to subvert facial recognition systems by inducing misclassification under diverse real-world environmental conditions. Unlike their digital counterparts, physical domain attacks must account for numerous environmental variables including illumination fluctuations, perspective variations, camera resolution limitations, and other physical constraints, rendering these attacks both more universally applicable and substantially more challenging to mitigate. Research has demonstrated that precisely calibrated light projection techniques can significantly impair facial recognition systems by disrupting the accurate extraction of biometric features. In a notable contribution to this field, C. Hu et al. introduced AdvIB, a sophisticated black-box physical attack methodology that compromises thermal infrared detection capabilities through the strategic optimization of physical parameters specifically the positioning, angular orientation, dimensional properties, and thermal intensity gradients of heat-generating and heat-absorbing elements discreetly embedded within subjects' attire [8]. The study begins by modeling the infrared block by defining four key physical parameters: position, pixel value, length, and angle, which are combined to form a set of infrared block configurations $\theta = \{B_1, B_2, \ldots, B_k\}$, where each block $B_1$ Contains the above parameters.Subsequently, within the digital environment, the parameters of the infrared blocks are optimized using the Differential Evolution (DE) algorithm. A set of initial parameter configurations (population) is randomly generated, with each individual representing a possible combination of blocks. Three distinct individuals are randomly selected from the population, and new parameters are generated based on the vector differences. As shown in Equation 7:

$$\theta_{new} = \theta_{base} + \text{mutation\_rate} \times (\theta_{individual1} - \theta_{individual2}) \quad (7)$$

The methodology uses stochastic parameter recombination to integrate parental and mutated variables probabilistically, preserving effective features. An iterative evaluation process assesses the adversarial efficacy of superior parameters for each subject until convergence. The Expectation over Transformation (EOT) framework simulates environmental complexities such as viewpoint and illumination changes to ensure the generated adversarial samples remain effective in natural settings. This optimization process produces a parameter set for generating digital adversarial samples. For physical implementation, the digital results are translated into tangible apparatus. Thermal elements (heating to 53°C) and cooling components (maintained at 24°C) simulate elevated and diminished infrared radiation, respectively. These elements are precisely positioned within the subject's garments across anterior, lateral, and posterior surfaces, with their angular orientation calibrated to ensure effectiveness across multiple perspectives. Figure 5 validates the attack effect in a real-world environment. The attack effect slightly decreases with increasing distance (lowest ASR of 40.8% at 8.6 meters), yet the overall success rate remains high. Multi-angle attacks are especially effective on the side and back.

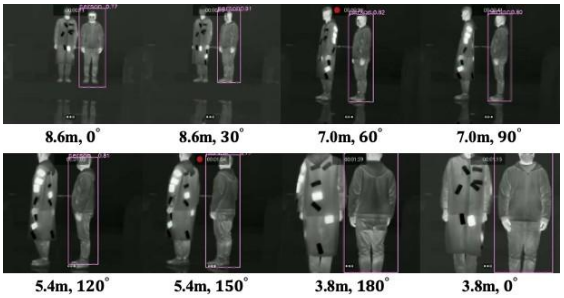


**Fig. 5** Demonstrating the effect of AdvIB attacks at different distances and angles [8]

### *3.2.2 Digital Domain Attacks*

It is entirely based on algorithms to manipulate images at the pixel level to realize end-to-end counterattacks. Typical methods, such as projected gradient descent (PGD), include the following process: take the original correctly classified samples as the starting point, set the number of iterations, the learning rate, and the perturbation range; compute the input gradient through cross-entropy loss, adjust the pixels along the gradient to maximize the probability of the model misclassification, and iteratively update it until the perturbation constraints are satisfied [9]. PGD not only generates strong adversarial samples, but also enhances the model through adversarial training robustness, and at the same time verifies the migratory nature of gradient optimization in black-box attacks, which becomes the core technique of digital domain attacks. The samples are updated according to the gradient direction and learning rate. As shown in Equation 8:

$$\mathrm{x_a dv} = \mathrm{x_a dv} + \eta * \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{L}(\mathrm{x_a dv}, \mathrm{y})) \tag{8}$$

where $x_a dv$ is the confrontation sample,Lis the loss function，y it's a real label，The sign function takes the sign of the gradient to ensure that the update is performed under the L∞ paradigm. Projection operation: project the updated sample back into the ε-neighbourhood of the original sample, Ensure that the difference between the confrontation sample and the original sample does not exceed the preset perturbation range. This step is achieved by the cropping operation, For example, under the L∞ paradigm, each pixel value of the sample is restricted to the range [-ε, +ε] of the original pixel value. When the number of iterations reaches a preset T, the iteration is stopped and the final adversarial sample is obtained. Loss function convergence: in some cases, if the loss function value changes less than a certain threshold in several consecutive iterations, the iteration can also be terminated early, and it is considered that a locally optimal adversarial sample has been found. Figure 6 below intuitively reveals the role of adversarial training in reshaping the internal structure of the model by visualising the parameter distributions, and at the same time provides a hierarchical analytical perspective for improving the defence strategy.

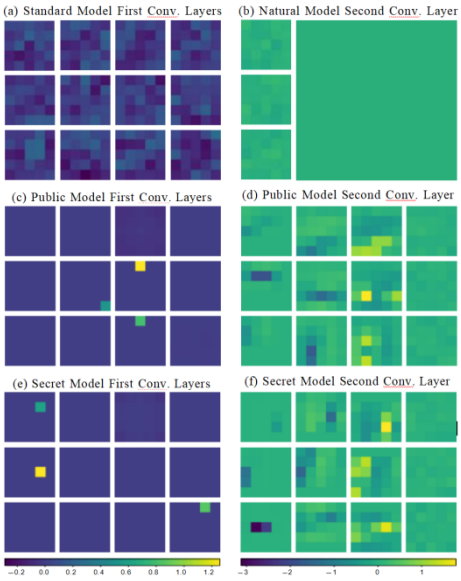


**Fig. 6** Displays parameter differences between the standard and adversarial training models at convolutional layer [9]

The ZOO attack by Chen et al. is a computationally efficient black - box method that generates adversarial samples by interrogating the target model's confidence outputs [10]. It innovates by formulating an objective function based on the model's confidence differentials, approximating gradients via finite difference methods, and combining stochastic coordinate descent with importance sampling to modify key pixels, reducing computational complexity. For high - dimensional models and large datasets like ImageNet, it balances efficacy and efficiency using techniques such as spatial downscaling and progressive dimensional expansion. Experiments show it effectively generates adversarial samples in complex scenarios and improves efficiency through multi - technique combinations. Figure 7 illustrates that the technique combination (red curve) achieves the best loss decrease, confirming its effectiveness in enhancing attack performance.

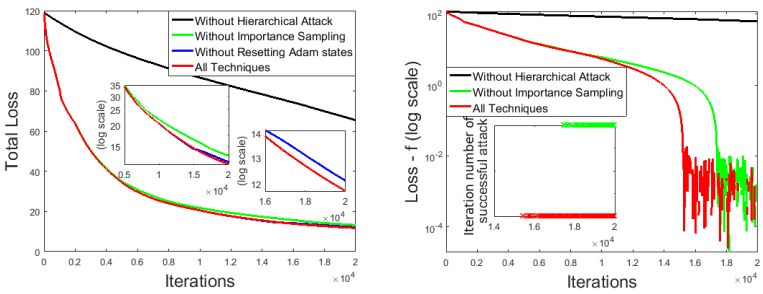


**Fig. 7** Curve of total loss with the number of iterations [10]

White-box attacks, as studied by Carlini and Wagner [4], involve attackers using gradient optimization to generate adversarial samples with full knowledge of the target model's structure, parameters, and gradients. Starting from original or slightly perturbed samples, the attack balances stealth and strength by designing a loss function (targeted

attack maximizes target category probability; non-targeted attack minimizes true category probability) and applying L2/L∞ constraints on perturbation magnitude. Algorithms like gradient descent iteratively adjust pixels to generate perturbations until termination conditions are met. Post-attack validation adjusts optimization strategies based on model misclassification and confidence changes. Figure 8 shows that on MNIST, some defenses can detect adversarial samples, but on CIFAR-10, they are harder to detect and most defenses are less effective.


**Fig. 8** Adversarial examples on the MNIST and CIFAR datasets for each defence studied [4]

# 4 Future Challenges

## 4.1 Improved Generalisation of Detection Methods

Current detection methodologies frequently suffer from excessive specialization toward particular datasets or specific falsification techniques, consequently exhibiting limited generalizability across diverse contexts. Promising avenues for future research encompass the development of cross-modal detection frameworks that integrate multimodal feature representations—including visual, acoustic, and behavioral signatures—to enhance discriminative capabilities. Additionally, the construction of more comprehensive training corpora and the strategic implementation of data augmentation techniques alongside synthetic data generation represent crucial strategies for expanding the effective training distribution, thereby substantially improving model robustness. These approaches collectively address the fundamental challenge of developing detection systems capable of maintaining high performance across heterogeneous, previously unencountered falsification methodologies and contextual environments.

## 4.2 Optimisation of Real-Time Detection in Low Resource Environments

Existing deep learning detection models usually have high computational complexity and

are difficult to be deployed in real-time on mobile or edge devices. Future research should focus on lightweight neural network design, such as efficient detection networks based on Transformer, efficient feature extraction and compression techniques to reduce model computational overhead and improve detection speed.

### 4.3 Detection and Defence of Physical Attack Scenarios

Most of the current detection methods mainly target adversarial attacks in digital space, while in the real world, attackers may use printed photos, screen flipping, etc. to bypass detection. Therefore, future research directions include constructing a framework for physical world adversarial attack detection, identifying forged samples using multi-angle sampling and infrared sensing, and combining 3D face recognition technology to enhance the defence capability through depth information.

### 4.4 Harmonised Assessment System Construction

Traditional attack success rates (ASR), though widely used, are static and cannot reflect defense degradation in sustained attacks. ForgeryNet's dynamic ASR (D-ASR) adds a time dimension (a 15% ASR rise after 24 hours) to better reflect real-world conditions. But current ASR evaluations mostly focus on single - type attacks (like FGSM or AdvHat), lacking coverage of hybrid digital - physical attacks. Komkov and Pautov note that physical - patch ASR needs assessment with factors like lighting and angles considered. SSIM often misleads by ignoring semantic distortion. Reference optimizes semantic consistency scores (SCS) via multimodal feature similarity (CLIP embeddings) to quantify the semantic fidelity of defense effects.

## 5 Conclusion

In recent years, although the wide application of deep learning in the field of face recognition has promoted technological innovation, its vulnerability to adversarial samples has exposed serious security risks. Studies have shown that by injecting tiny perturbations in the image that are indistinguishable to the human eye such as FGSM and C&W attacks, or by using physical patches such as antagonistic glasses and AdvHat, the model can be easily induced to misjudge and threaten critical scenarios such as security.

For such attacks, the detection technology gradually evolves from traditional machine learning PCA+SVM to deep learning, end-to-end network SmartBox framework. High-precision detection is achieved through multi-scale feature fusion (AUC 0.93 for the LFW dataset), while Vision Transformer (ViT-Detect) excels in cross-domain testing by virtue of its long-range dependency modelling capability, but its computational overhead constrains edge deployment.

Adversarial perturbation removal techniques, such as residual generation networks, can eliminate some noise but still need robustness improvement against strong adversarial samples (e.g., EAD attacks). Physical attacks (e.g., stickers, make - up perturbations) have high real -world success rates, showing models' sensitivity to local perturbations. Existing evaluation systems have shortcomings, such as ASR ignoring dynamic attenuation. The introduction of Dynamic ASR (D - ASR) and Semantic Consistency Score (SCS) offers new multi - dimensional assessment directions

## References

1. Sharif, M., Bhagavatula, S., Bauer, L., et al.: 'Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition', in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 2016, pp. 1528-1540
2. Komkov, S., Petiushko, A.: 'AdvHat: Real-World Adversarial Attack on ArcFace Face ID System', in 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 819-826
3. Goodfellow, I. J., Shlens, J., Szegedy, C.: 'Explaining and harnessing adversarial examples', arXiv preprint arXiv:1412.6572, 2014
4. Carlini, N., Wagner, D.: 'Towards Evaluating the Robustness of Neural Networks', in 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 39-57
5. Chen, P. Y., Sharma, Y., Zhang, H., et al.: 'EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples', in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018
6. Eykholt, K., Evtimov, I., Fernandes, E., et al.: 'Robust Physical-World Attacks on Deep Learning Models', in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1625-1634
7. Kurakin, A., Goodfellow, I. J., Bengio, S.: 'Adversarial examples in the physical world', arXiv preprint arXiv:1607.02533, 2016
8. Hu, C., Shi, T., Jiang, W., et al.: 'Adversarial Infrared Blocks: A Black-box Attack to Thermal Infrared Detectors at Multiple Angles in Physical World', in Proceedings of the ACM Conference on Computer and Communications Security (CCS), Washington, DC, USA, 2017, pp. 1-10
9. Madry, A., Makelov, A., Schmidt, L., et al.: 'Towards Deep Learning Models Resistant to Adversarial Attacks', in Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), 2018
10. Chen, P. Y., Zhang, H., Sharma, Y., et al.: 'ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models', in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17), New York, NY, USA, 2017, pp. 15-26