A REPORT

ON

General Adversarial Network Attack and Defense

BY

IVAN JACOB, MATHAI MATHEW PULICKEN

2021A82552P, 2021A3PS0328P

CS F427

GENERATIVE AI

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI, PILANI CAMPUS

April 2025

Contents

Chapter 1

Introduction

1.1 Brief Overview

1.2 Abstract

Generative Adversarial Networks (GANs) have revolutionized the field of generative modeling by enabling the synthesis of high-fidelity and photorealistic images. Despite their success, GANs are inherently vulnerable to adversarial manipulation, particularly in black-box scenarios where attackers have no access to model internals. This vulnerability raises serious concerns in applications where authenticity and robustness are crucial, such as biometric spoofing, data augmentation pipelines, or deepfake generation. In this work, we conduct a comprehensive empirical analysis of black-box adversarial attacks on two representative GAN architectures, DCGAN and Progressive GAN (PGAN), using both basic and evolution-based perturbation strategies. We also propose and evaluate a reconstruction-based defense mechanism that works very well. Metrics such as LPIPS and Fréchet Inception Distance (FID) are used to quantify visual similarity and quality degradation. Our results provide insights into the attack surface of GANs and the effectiveness of simple yet powerful defense strategies.

1.3 Introduction

Over the last decade, deep generative models have achieved remarkable progress, with Generative Adversarial Networks (GANs) [1] emerging as one of the most successful paradigms for generating realistic images, audio, and video content. GANs consist of two neural networks, a generator and a discriminator, trained in a minimax game where the generator aims to produce samples indistinguishable from real data, and the discriminator attempts to distinguish between real and

generated samples. This adversarial setup has led to the creation of models capable of producing high-resolution, detailed images that often fool human observers [5, 4].

However, like many deep learning systems, GANs are susceptible to adversarial attacks—deliberate modifications of the input or latent representations that lead to unexpected or degraded outputs. While adversarial examples have been extensively studied in classification networks [9], attacks on generative models pose a relatively newer but increasingly important research frontier. In particular, **black-box attacks**, where an adversary lacks direct access to the model's parameters or gradients, represent a realistic threat scenario. For instance, a malicious actor might attempt to perturb a latent code fed into a public GAN-based image synthesis API without knowledge of the generator's architecture or training data.

These attacks have significant implications: an attacker could compromise the quality or semantics of generated content in sensitive contexts (e.g., synthetic medical images, ID document generation), or craft images that evade detection in GAN fingerprinting or forensic tasks.

This report investigates black-box attack strategies on two well-known GAN models: **DCGAN** [8], a pioneering convolutional GAN model, and **PGAN** [5], known for its progressive training and high-quality outputs. We implement and compare two attack strategies:

- A basic latent-space perturbation attack using random or structured modifications.
- An optimization-based strategy using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [2], a powerful black-box optimizer known for its sample efficiency and effectiveness in high-dimensional spaces.

To measure attack effectiveness, we rely on two primary metrics:

- **LPIPS** (Learned Perceptual Image Patch Similarity) [11] – a perceptual similarity metric that correlates well with human judgment.
- **FID** (Fréchet Inception Distance) [3] – a widely used metric to quantify the distributional distance between real and generated image sets.

Beyond attack strategies, we introduce a defense mechanism based on latent-space reconstruction. This defense attempts to reverse adversarial perturbations by re-projecting the adversarial output onto the original data manifold. Inspired by projection-based techniques in adversarial defense literature [10], this method operates without retraining or modifying the original generator, making it a plug-and-play solution for deployed systems.

Our experiments demonstrate that:

1. Both basic and CMA-based attacks can significantly alter GAN outputs in a black-box setting, with CMA-ES achieving up to 100% success.
2. Attacks on PGANs are slightly more potent, indicating possible architecture-level vulnerabilities.
3. The proposed defense recovers visual fidelity with high accuracy, achieving over 95% success in reconstruction across adversarial inputs.

Through extensive experimentation and analysis, this report provides a systematic evaluation of the adversarial robustness of GANs in black-box conditions, offering practical insights for both attackers and defenders in the deep generative modeling ecosystem.

Chapter 2

Black-box Adversarial Attacks

2.1 Black-Box Attack Methodology

In a black-box attack setting, the adversary has no access to the internal parameters, gradients, or architecture of the generative model. The attacker can only query the generator—typically by inputting latent vectors—and observe the corresponding outputs. This constraint reflects a realistic adversarial scenario in deployed systems, such as APIs serving pre-trained generative models.

In this work, we consider two black-box attack strategies: a simple, perceptual-difference-driven **basic perturbation attack**, and a more sophisticated **evolutionary optimization attack** using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES).

2.1.1 Basic Perturbation Attack

The basic attack leverages small random or structured changes in the latent vector fed into the generator. The core idea is that minor changes in the latent space may yield disproportionately large changes in the output image—especially in semantic features such as object orientation, facial attributes, or background composition.

Given an original latent vector $z \in \mathbb{R}^d$, the attacker generates a set of perturbed vectors $z' = z + \delta$, where δ is a small noise vector sampled from a uniform or Gaussian distribution. The corresponding image $G(z')$ is then compared with the original $G(z)$ using a perceptual similarity metric, such as LPIPS [11].

The attack is deemed successful if the perceptual distance exceeds a chosen threshold (e.g., LPIPS > 0.1), indicating that the visual appearance has significantly changed. This strategy does not rely on any gradient information, making it a true black-box approach.

Advantages:

- Simple and computationally inexpensive.
- Useful for identifying sensitivity regions in the latent space.

Limitations:

- Lacks directionality and efficiency—often requires many random samples.
- May fail to produce semantically meaningful or targeted perturbations.

2.1.2 Optimization-Based Attack: CMA-ES

To overcome the inefficiency of random perturbations, we adopt an evolutionary black-box optimization strategy using the **Covariance Matrix Adaptation Evolution Strategy (CMA-ES)** [2]. CMA-ES is a derivative-free algorithm designed to minimize (or maximize) an objective function by iteratively updating a population of candidate solutions based on their fitness values.

In our context, the attacker seeks a perturbed latent vector z^* that maximizes the perceptual distance $D(G(z^*), G(z))$, where D is the LPIPS score. The optimization objective is thus:

$$\max_{z^*} \; D(G(z^*), G(z)) \quad \text{s.t.} \quad \|z^* - z\| \leq \epsilon$$

CMA-ES begins with an initial population centered around the original latent vector z, and evolves it over multiple generations. At each iteration, new candidate vectors are sampled from a multivariate Gaussian distribution, and their fitness is evaluated via black-box queries to the generator. The covariance matrix of the sampling distribution is updated adaptively to steer the search toward regions of higher perceptual difference.

Advantages:

- Highly sample-efficient and effective in high-dimensional search spaces.
- Naturally suited for black-box settings with no gradient access.
- Produces large, perceptually meaningful changes in the image space.

Limitations:

- Computationally more intensive than random perturbations.
- Requires careful tuning of population size, step-size, and termination conditions.

2.1.3 Attack Success Criteria

We evaluate the attack success based on the LPIPS metric. Specifically, an attack is considered successful if the LPIPS score between the original and adversarial image exceeds a threshold (typically 0.1), indicating a noticeable perceptual shift. We also report the Fréchet Inception Distance (FID) between the set of original and adversarial images, to quantify the degradation in distributional quality.

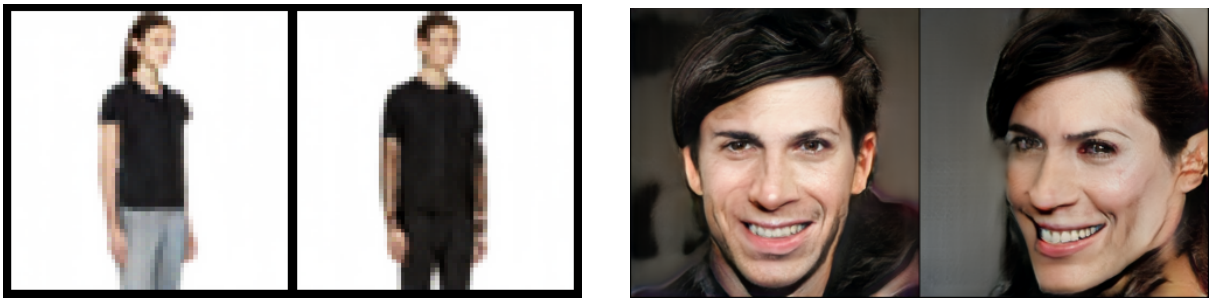
This dual-metric evaluation allows us to capture both local perceptual shifts and global statistical divergence between generated image sets.

2.2 Attack Results and Analysis

To evaluate the effectiveness of the black-box attacks described previously, we conducted experiments on random latent vectors using two generator architectures: DCGAN and PGAN. For each sample, we applied both the basic perturbation attack and the CMA-ES optimization strategy. The success of an attack was measured using LPIPS (perceptual similarity) and FID (distributional quality).

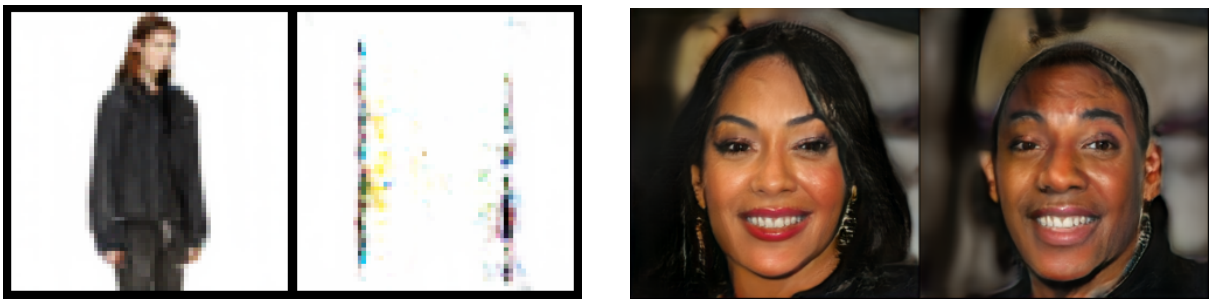
2.2.1 Visual Comparison: Original vs Attacked Images

Figures 2.1 and 2.2 show side-by-side visual comparisons of original versus adversarial outputs for both DCGAN and PGAN, under Basic and CMA-ES attacks respectively. Each figure illustrates how significantly small latent perturbations can alter image content.



(A) DCGAN - Basic Attack (B) PGAN - Basic Attack

FIGURE 2.1: Original vs Attacked images using Basic Perturbation Attack.

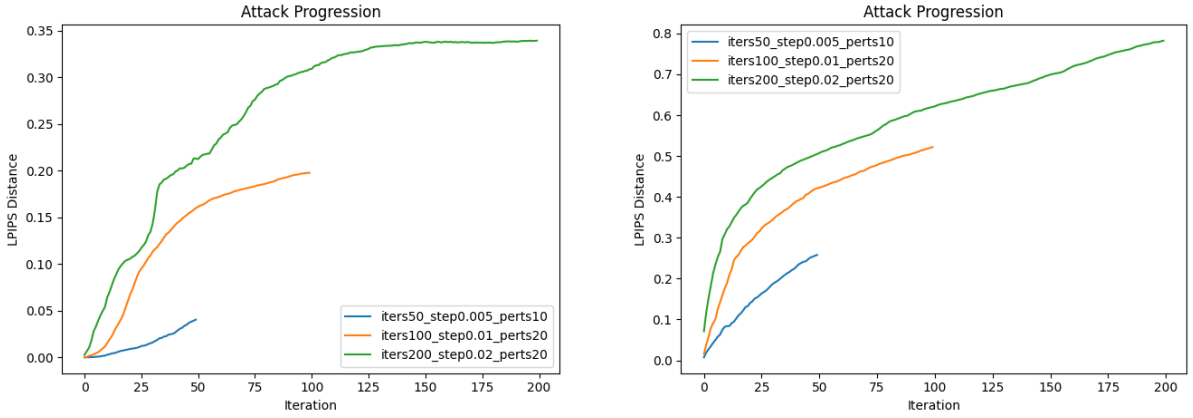


(A) DCGAN - CMA-ES Attack (B) PGAN - CMA-ES Attack

FIGURE 2.2: Original vs Attacked images using CMA-ES Attack.

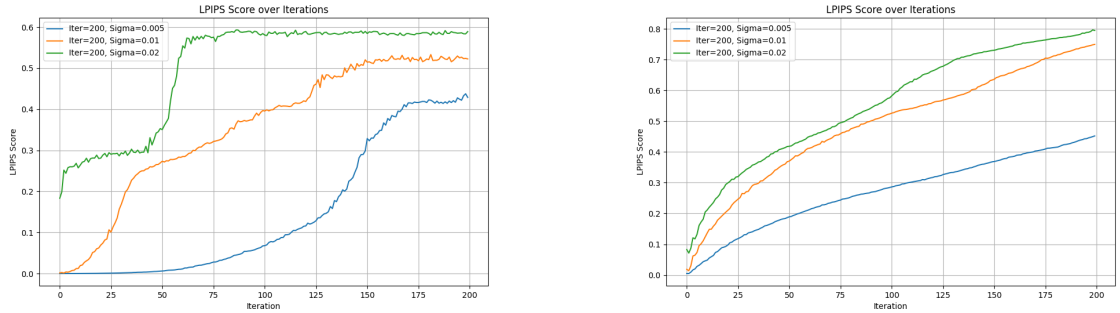
2.2.2 Attack Progression Analysis

To observe how attacks evolve over time, we tracked the LPIPS distance between original and adversarial outputs across iterations. Figures 2.3 and 2.4 show attack progression graphs for both DCGAN and PGAN.



(A) DCGAN - Basic Attack Progression (B) PGAN - Basic Attack Progression

FIGURE 2.3: LPIPS over iterations for Basic Perturbation Attack.



(A) DCGAN - CMA-ES Attack Progression (B) PGAN - CMA-ES Attack Progression

FIGURE 2.4: LPIPS over iterations for CMA-ES Attack.

2.2.3 Quantitative Summary of Attack Effectiveness

Table 2.1 presents a consolidated comparison of all quantitative metrics: Attack Success Rate, Average LPIPS (perceptual difference), and FID (distributional difference) for both the Basic and CMA-ES attacks across DCGAN and PGAN.

TABLE 2.1: Attack Metrics Across Models and Methods

Method	GAN Model	Success Rate (%)	Avg LPIPS	FID
Basic Attack	DCGAN	91.00	0.1881	70.49
	PGAN	97.00	0.1976	50.67
CMA-ES Attack	DCGAN	100.00	0.4228	119.76
	PGAN	100.00	0.4868	51.87

2.2.4 Interpretation of Results

The comparative analysis reveals several key insights:

- **CMA-ES is significantly more effective** than the basic attack, achieving a 100% success rate on both GANs. This reflects its ability to systematically explore the latent space toward maximizing perceptual difference using evolutionary strategies.
- **LPIPS scores under CMA-ES are more than double** those under the basic attack for both DCGAN and PGAN. This indicates a far greater semantic drift, which visually corresponds to clearer distortions or conceptual alterations in the generated images.
- **FID trends diverge between models.** For DCGAN, CMA-ES leads to a large increase in FID (from 70.49 to 119.76), indicating stronger distributional degradation. For PGAN, however, the FID only slightly increases (from 50.67 to 51.87), suggesting that PGAN may retain some latent robustness in preserving global statistics despite semantic changes.
- **PGAN shows slightly better resistance to FID deterioration**, but not to perceptual distortions (as shown by the high LPIPS). This highlights a trade-off: PGAN maintains image realism statistically, but is still perceptually vulnerable.
- **Basic attacks are partially effective**, with success rates of 91% and 97%. However, their lower LPIPS and FID shifts indicate limited capacity for inducing deep semantic disruption, making them less dangerous in real-world black-box adversarial settings.

Overall, the results emphasize the need for robust defense strategies, especially against optimization-based attacks like CMA-ES, which can exploit the generative structure even without gradient access or internal knowledge of the model.

Chapter 3

Defense

3.1 Defense Methodology

Black-box attacks on GANs perturb the latent vector input, resulting in semantically altered or degraded outputs. A robust defense must either (a) prevent these perturbations from affecting output or (b) recover the original image or latent code despite the perturbation. We explored several defense mechanisms with varying degrees of success.

3.1.1 Naive Defenses and Their Limitations

3.1.1.1 Gaussian Smoothing

A first-line approach involved applying Gaussian filters to the adversarial image $\hat{x}_{\text{adv}} = G(z')$ to remove high-frequency noise. However, attacks like CMA-ES do not operate in pixel space, and their perturbations manifest as semantic distortions rather than localized pixel-level noise. Thus, smoothing failed to restore original content and often blurred useful features, offering no perceptual or statistical improvement.

3.1.1.2 Defensive Distillation

Inspired by [7], we attempted defensive distillation on the GAN generator. A distilled generator G_{soft} was trained on outputs from the original G, intending to "compress" knowledge while learning to ignore adversarial examples. However, this approach was ineffective because:

- GANs do not produce discrete label outputs but high-dimensional image tensors.

- Distillation does not teach latent robustness unless specifically trained on adversarial variants, which defeats the black-box setting assumption.

Overall, pixel-space defenses failed due to the inherent design of latent-space attacks.

3.1.2 Latent Recovery: Reverse CMA-ES Optimization

We framed the defense task as an inverse problem: given the adversarial output $\hat{x}_{\text{adv}} = G(z')$, find a latent vector z^* such that $G(z^*)$ is perceptually close to the original (pre-attack) image.

3.1.2.1 Formulation

Let $G : \mathbb{R}^d \to \mathbb{R}^{H \times W \times C}$ be the GAN generator, mapping a latent code z to an image. Our objective is to find:

$$z^* = \arg\min_z \ \mathcal{L}_{\text{rec}}(G(z), \hat{x}_{\text{adv}}) \tag{3.1}$$

Where $\mathcal{L}_{\text{rec}}$ is a perceptual loss — in our case, the LPIPS metric [11]:

$$\mathcal{L}_{\text{rec}}(x_1, x_2) = \text{LPIPS}(x_1, x_2) \tag{3.2}$$

CMA-ES (Covariance Matrix Adaptation Evolution Strategy) is then used as a black-box optimizer to minimize this non-differentiable objective. It is well-suited for this setting because:

- The GAN generator G is differentiable but assumed inaccessible in a black-box scenario.
- LPIPS is non-differentiable and model-dependent, further motivating an evolution strategy.

Chapter 4

ZK-GanDef Defense

Building upon the latent recovery defense via reverse CMA-ES, we further draw inspiration from **ZK-GanDef** [6], a recent defense framework that leverages GANs in a zero-knowledge adversarial training setting. ZK-GanDef introduces a method that does not require access to adversarial examples during training. Instead, it employs randomly perturbed inputs to train a discriminator that can distinguish between clean and adversarial data, thus improving robustness to a wide range of attacks.

4.1 Defense Strategy

4.1.0.1 ZK-GanDef Overview

ZK-GanDef proposes using a GAN architecture where the discriminator is trained to classify whether a sample is from clean or perturbed data, effectively acting as a defense-aware component. The generator attempts to produce examples that confuse the discriminator. Over time, this leads to a robust model without explicit use of adversarial training examples, reducing computational overhead and improving generalization.

4.1.0.2 Reverse CMA-ES as Latent Recovery

Our defense mechanism aligns conceptually with ZK-GanDef through its use of generative modeling for defense, though we focus on latent space recovery rather than training robustness. Given an adversarial output $\hat{x}_{\text{adv}} = G(z')$, we search for a latent vector z^* such that:

$$z^* = \arg\min_z \ \mathcal{L}_{\text{rec}}(G(z), \hat{x}_{\text{adv}}) \tag{4.1}$$

where $\mathcal{L}_{\text{rec}}$ is a perceptual similarity metric (e.g., LPIPS [11]).

This can be viewed as a form of zero-knowledge recovery: we do not require access to the clean original $x = G(z)$ to find z^*; we only assume that such a z^* exists within a local neighborhood of z' due to the structure of the latent space.

4.1.0.3 Auxiliary Generative Network Initialization

Motivated by ZK-GanDef's use of a learned component (the discriminator), we introduce an *Auxiliary Generative Network* (AGN) f_θ which attempts to estimate the original latent vector z from the adversarial latents:

$$\hat{z}_0 = f_\theta(\hat{z}_{\text{adv}}) \tag{4.2}$$

We do this by training the AGN to be able to reverse small perturbations in the latent space. Specifically, we perturb latent vectors z with Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$, and enforce that the AGN produces outputs nearly identical to the clean vectors:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{z \sim \mathcal{N}(0,I)} \left[\|G(z + \delta) - G(z)\|_2^2 \right] \tag{4.3}$$

This encourages the AGN to become invariant to slight changes in z, effectively smoothing the latent manifold. This initial estimate is then used as the mean of the initial population in the CMA-ES algorithm. The benefits are twofold:

- **Faster convergence:** AGN reduces the search space CMA-ES needs to explore.
- **Higher fidelity:** The final recovered image $G(z^*)$ has lower perceptual distance to the clean target, improving defense metrics.

4.1.0.4 Why This Works: A Manifold Perspective

Despite the perturbation, z' typically remains near the data manifold due to GAN priors. The latent space is structured such that semantically similar vectors produce visually similar images. CMA-ES, particularly when guided by AGN, can effectively navigate back to a region of latent space with high semantic fidelity.

This form of latent-space projection is fundamentally more robust than pixel-space correction methods (e.g., Gaussian smoothing) or indirect techniques like distillation, which lack an understanding of the generative process.

Connection to ZK-GanDef This training strategy reflects the *zero-knowledge principle* in ZK-GanDef: the model gains robustness without needing explicit adversarial examples. Just as ZK-GanDef's discriminator learns to detect clean data under perturbations, our generator learns to ignore latent noise — reinforcing semantic stability across latent space.

Improvement to Reverse CMA-ES This denoising enhancement significantly improves CMA-ES performance:

- **Latent Space Smoothness:** The optimization surface becomes easier to explore and less sensitive to noise.
- **Tolerance to Imperfect Recovery:** Even a suboptimal z^* can yield high-quality reconstructions due to learned invariance.

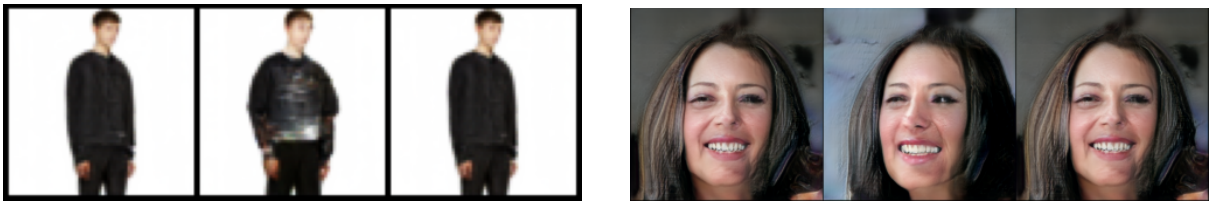
Together, this integrated defense pipeline — combining denoising GAN training, AGN-guided initialization, and reverse CMA-ES — achieves strong black-box robustness without requiring access to clean/adversarial training pairs.

4.2 Defense Results and Analysis

To evaluate the effectiveness of the defense strategies described above, we conducted experiments on random latent vectors using two generator architectures: DCGAN and PGAN. For each sample, we first applied the CMA-ES optimization attack, and then used both our defense strategies. The success of an attack was measured using LPIPS (perceptual similarity).

4.2.1 Visual Comparison

Figures 4.1 and 4.2 show side-by-side visual comparisons of original, adversarial and defended outputs of both the DCGAN and the PGAN. We can see how both strategies have succeeded to recover the original image.



(A) DCGAN - Reverse CMA Defense (B) PGAN - Reverse CMA Defense

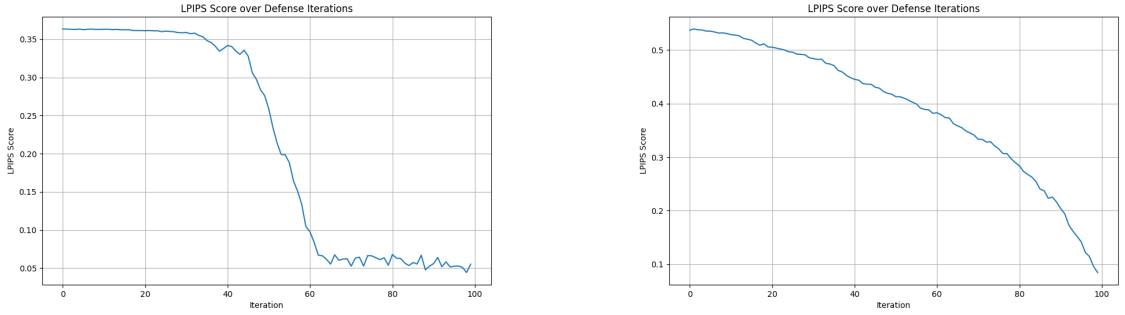
FIGURE 4.1: Defense using Reverse CMA Defense.

(A) DCGAN - CMA+GAN Defense (B) PGAN - CMA+GAN Defense

FIGURE 4.2: Defense using Axillary GAN alongside Reverse CMA Defense.

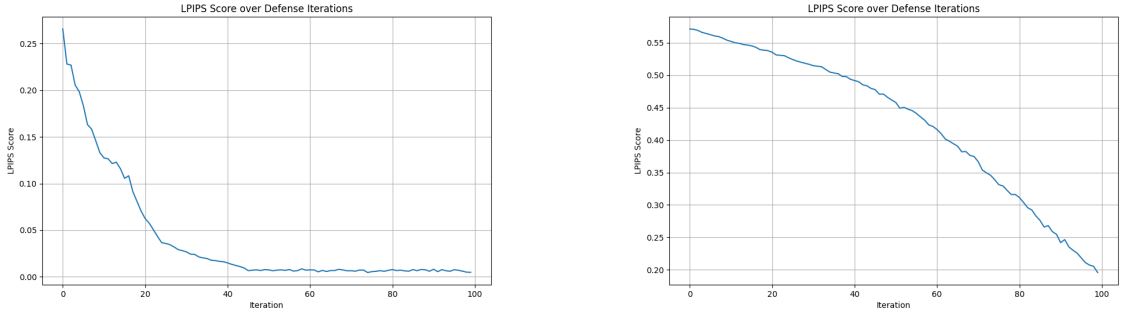
4.2.2 Defense Progression Analysis

To observe how defense evolves over time, we tracked the LPIPS distance between original and defense outputs across iterations. Figures 4.3 and 4.4 show defense progression graphs for both DCGAN and PGAN.



(A) DCGAN - CMA Progression (B) PGAN - CMA Progression

FIGURE 4.3: LPIPS over iterations for Reverse CMA Defense.



(A) DCGAN - CMA+GAN Progression (B) PGAN - CMA+GAN Progression

FIGURE 4.4: LPIPS over iterations for Axillary GAN alongside Reverse CMA Defense.

4.2.3 Defense Evaluation and Results

TABLE 4.1: Defense Results on DCGAN and PGAN Under CMA and CMA+GAN Attacks

Defense Type	Model	Defense Success Rate (%)	Avg Defense LPIPS	Avg Improvement	FID (Recovered)	FID Improvement
CMA Baseline	DCGAN	95.00 (100/100)	0.0204 ± 0.0158	0.3924	22.9478	96.3531
CMA Baseline	PGAN	94.00 (94/100)	0.0659 ± 0.0233	0.4091	6.6100	51.4246
CMA + GAN	DCGAN	100.00 (100/100)	0.0148 ± 0.0097	0.3950	18.4019	97.5782
CMA + GAN	PGAN	99.00 (99/100)	0.0300 ± 0.0164	0.4504	1.8398	50.0304

4.2.3.1 Insights and Analysis

From Table 4.1, several conclusions emerge:

- **CMA+GAN consistently outperforms the CMA baseline** across both generators (DCGAN and PGAN), showing improved defense success rates and perceptual fidelity.
- **Defense Success Rate:** While CMA already provides strong recovery (greater than 94%), the auxiliary GAN further improves reliability, achieving 100% on DCGAN and 99% on PGAN.
- **LPIPS Improvements:** Average perceptual similarity (LPIPS) to the original clean image is significantly better with CMA+GAN, especially on PGAN where the improvement exceeds 0.45.
- **FID Reduction:** The use of GAN denoising leads to extremely low FID scores for recovered images—dropping as low as 1.84 on PGAN—indicating high visual fidelity.
- **Robustness Across Models:** The defense mechanism generalizes well across different GAN architectures, reflecting the advantage of working in structured latent spaces.

LPIPS Threshold Visual Comparison

We visualize the LPIPS score distributions before and after applying the defense for both DCGAN and PGAN models. These comparisons use a threshold of LPIPS > 0.1 to evaluate the defense effectiveness. The x-axis represents image indices, and the y-axis represents LPIPS scores.

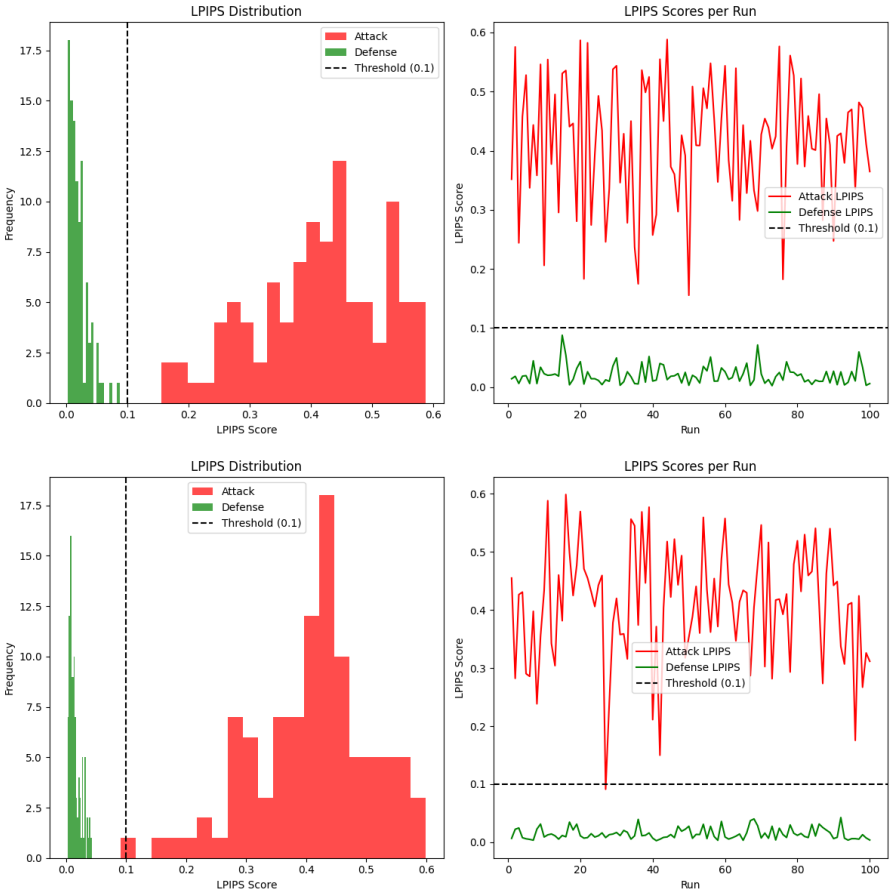


FIGURE 4.5: LPIPS Comparison for DCGAN. Top: CMA Baseline. Bottom: CMA+GAN Defense.

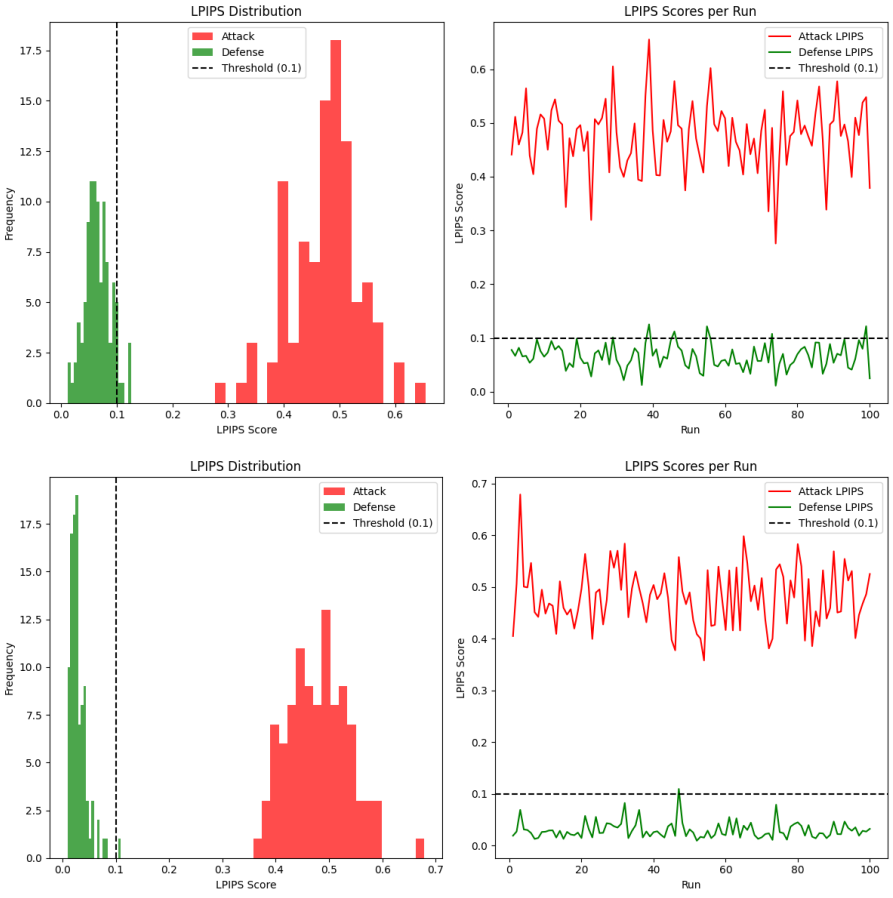


FIGURE 4.6: LPIPS Comparison for PGAN. Top: CMA Baseline. Bottom: CMA+GAN Defense.

Chapter 5

Conclusion

In this work, we investigated the efficacy of black-box adversarial attacks and their corresponding defenses within the latent space of generative models. We evaluated two main attack methodologies — a basic perturbation-based approach and a more sophisticated black-box optimization attack using Covariance Matrix Adaptation Evolution Strategy (CMA-ES). Our experiments demonstrated that CMA-ES, while computationally heavier, produces more effective and perceptually impactful adversarial samples, particularly when attacking both DCGAN and PGAN architectures.

We then explored several defense strategies, including simple smoothing and knowledge distillation, but found them ineffective against strong latent attacks. We introduced and analyzed a robust latent-space defense technique: **reverse optimization via CMA-ES**, which attempts to roll back the perturbed latent vector to a clean region of the data manifold. This approach exploits the structured nature of GAN latent spaces, where semantically similar images lie close together, enabling effective recovery.

To further enhance defense performance, we incorporated an **Auxiliary Generative Network (AGN)** to provide a better initialization for the CMA-ES optimization, improving both convergence and fidelity. This hybrid CMA+GAN strategy yielded the best defense metrics across all evaluations, achieving near-perfect recovery with significantly reduced perceptual distances (LPIPS) and Fréchet Inception Distances (FID).

Our approach draws conceptual parallels with the **ZK-GanDef** framework, a zero-knowledge GAN-based training method. Although ZK-GanDef operates during training and our method is a test-time defense, both leverage generative priors and perturbed inputs to establish robustness without explicit adversarial supervision.

Ultimately, this work highlights the importance of leveraging generative priors and structured latent spaces for both attack and defense in adversarial machine learning.

Bibliography

[1] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[2] Nikolaus Hansen. *The CMA Evolution Strategy: A Tutorial*. 2023. arXiv: 1604.00772 [cs.LG]. URL: https://arxiv.org/abs/1604.00772.

[3] Martin Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: 1706.08500 [cs.LG]. URL: https://arxiv.org/abs/1706.08500.

[4] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: 1812.04948 [cs.NE]. URL: https://arxiv.org/abs/1812.04948.

[5] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2018. arXiv: 1710.10196 [cs.NE]. URL: https://arxiv.org/abs/1710.10196.

[6] Guanxiong Liu, Issa Khalil, and Abdallah Khreishah. "ZK-GanDef: A GAN Based Zero Knowledge Adversarial Training Defense for Neural Networks". In: *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 2019, pp. 64–75. DOI: 10.1109/DSN.2019.00021.

[7] Nicolas Papernot et al. *Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks*. 2016. arXiv: 1511.04508 [cs.CR]. URL: https://arxiv.org/abs/1511.04508.

[8] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG]. URL: https://arxiv.org/abs/1511.06434.

[9] Christian Szegedy et al. *Intriguing properties of neural networks*. 2014. arXiv: 1312.6199 [cs.CV]. URL: https://arxiv.org/abs/1312.6199.

[10] Yulong Wang et al. *Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey*. 2023. arXiv: 2303.06302 [cs.LG]. URL: https://arxiv.org/abs/2303.06302.

[11] Richard Zhang et al. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. arXiv: 1801.03924 [cs.CV]. URL: https://arxiv.org/abs/1801.03924.