Biometric Leakage from Generative Models and Adversarial Iris Swapping for Spoofing Eye-based Authentication

Jan Michalak

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Masters of Science
in
Computer Science Applications

Brendan David-John, Chair
Bo Ji
Bimal Viswanath

May 13, 2025
Blacksburg, Virginia

Keywords: Iris Authentication, Digital Presentation Attacks, Generative Adversarial Networks (GANs), Diffusion Models, AR/VR Biometrics

Biometric Leakage from Generative Models and Adversarial Iris Swapping for Spoofing Eye-based Authentication

Jan Michalak

(ABSTRACT)

This thesis investigates the vulnerability of generative models trained on biometric data and explores digital spoofing attacks on iris-based authentication systems representative of AR/VR environments. We first explore how diffusion models trained on biometric data can memorize and leak iris images. Next, we evaluate the effectiveness of Cross-Attention GANs for iris-swapping attacks, demonstrating their ability to enable presentation attacks that spoof iris-recognition systems. Our experiments across several standard iris and VR datasets have an attack success rate of 100% within similar domains and generalize across domains with rates as high as 70%. Our findings highlight the need to consider vulnerabilities in biometric systems and strengthen defenses against digital presentation attacks produced by generative models.

Biometric Leakage from Generative Models and Adversarial Iris Swapping for Spoofing Eye-based Authentication

Jan Michalak

(GENERAL AUDIENCE ABSTRACT)

Most people are familiar with Face ID, which uses facial features to unlock phones or laptops. But in virtual and augmented reality (AR/VR) headsets, where faces are not always visible, devices rely on the iris to recognize users. In this thesis, we show that AI models trained on iris images can sometimes memorize and leak information about the people they were trained on. We also designed a new type of attack where a fake eye is created by swapping one person's iris onto another's eye. These generated irises can trick iris recognition systems into thinking the attacker is the target. Our results show near-perfect success when attacking within the same dataset, and strong success when crossing between different VR datasets. These findings suggest that while iris recognition holds promise for secure login in AR/VR devices, it also opens a new risk that AI-powered attacks could exploit.

Acknowledgments

I would first like to express my deepest appreciation to my advisor, Dr. Brendan David-John. Working under his guidance has been one of the most rewarding experiences within my academic career. His passion for research is truly inspiring, and his belief in me made all the difference, especially in times of self-doubt. Beyond teaching me how to do research, he showed me how to stay curious, think deeply, and made me excited about the process of discovery. The lessons he taught me will stay with me far beyond the completion of this thesis. I'll forever appreciate the countless hours he spent helping me grow both as a person and a scholar. I'd also like to thank my lab mates in the PrivateEye Lab for all of the many conversations, advice, and encouragement that made the most challenging days easier to navigate. Among them, I'd like to especially thank Anish Narkar for mentorship across different projects. His guidance and willingness to share his knowledge have helped me to grow as a researcher and have been invaluable to my time in my program. I'd also like to thank the friends that I've made along the way, Nicholas Kong, Sophia Stil, and James Weichert, for being a grounding presence throughout this process. Lastly, I'd like to thank my parents, Kasia and Pawel Michalak, for their constant wisdom, guidance, and unconditional support throughout my time at VT.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Background

Biometric authentication is becoming a familiar part of everyday technology, with iris recognition playing an increasingly important role. What sets iris biometrics apart is their remarkable uniqueness and long-term stability, making them a trusted standard for secure identification. While face and fingerprint recognition are common on mobile devices, iris-based systems are gaining traction on mixed reality (MR) platforms like the Microsoft HoloLens 2, Magic Leap 2, and Apple Vision Pro, where capturing a full facial image is not always possible.

Even though iris-based authentication is considered highly secure, it is not invulnerable. Presentation attacks — where an attacker tries to trick the system using fake biometric inputs — continue to pose a real challenge. In the past, most defenses have focused on physical attacks, such as printed iris images, textured contact lenses, or video replays. But the landscape is shifting. The rise of generative AI, especially models like Generative Adversarial Networks (GANs) and diffusion models, has opened the door to a new type of threat: digitally crafted biometric spoofs. These technologies can create synthetic but convincing iris patterns, raising new concerns about the ability to bypass even the most robust iris authentication systems.

1.2 Problem Statement

The primary challenge addressed in this thesis is the vulnerability of iris-based authentication systems, especially those embedded in AR/VR headsets, to digital presentation attacks enabled by generative models.

This thesis explores two key threat vectors:

1. The ability of diffusion models to memorize and regenerate real identities from training data, leading to biometric leakage.
2. The feasibility and robustness of iris-swapping attacks using cross-attention GAN architectures, especially across different domains.

1.3 Research Questions

This thesis investigates the following questions:

1. **RQ1: Can generative models trained on iris data leak memorized identities?**
   Does fine-tuning a diffusion model on iris datasets result in the reproduction of training samples that are biometrically valid under standard recognition systems?
2. **RQ2: Can a GAN-based iris-swapping method impersonate identities using known iris data?**
   Using a known target iris, can a Cross-Attention GAN model generate realistic swaps that match the target identity while maintaining natural eye appearance?
3. **RQ3: Do iris-swapping attacks generalize across AR/VR datasets and biometric pipelines?**

Can a trained swap model maintain spoofing effectiveness on datasets captured under AR/VR-specific conditions, such as variable lighting, pose, and sensor type?

1.4 Thesis Organization

This thesis is structured as follows:

- **Chapter 1: Introduction** outlines the motivation for this work, the core problem statement, and the three research questions that guide the study.
- **Chapter 2: Threat Model** defines the adversary's capabilities and assumptions, and describes the risk landscape for iris-based authentication in AR/VR environments.
- **Chapter 3: Background** reviews related work in iris biometrics, presentation attacks (both physical and digital), and vulnerabilities in machine learning models, particularly generative models.
- **Chapter 4: Methodology and Preliminary Experiments** introduces two attack pipelines: identity leakage via fine-tuned diffusion models (**RQ1**), and digital iris-swapping attacks using a Cross-Attention GAN (**RQ2 and RQ3**). For the leakage attack, we present both the methodology and results within this chapter, as the evaluation is self-contained
- **Chapter 5: Results** presents the results of swap-based attacks against iris recognizers across different domains and recognition settings. This chapter directly addresses **RQ2** and **RQ3**, analyzing spoofing success in white-box and black-box conditions, and generalization across datasets.
- **Chapter 6: Conclusions** discusses results, limitations, and avenues for future work.

Chapter 2

Threat Model

2.1 Context

Iris-based biometric authentication is gaining traction as a secure and seamless mode of identity verification, especially within emerging Virtual Reality (VR) and Augmented Reality (AR) devices. Devices such as the Apple Vision Pro, Microsoft HoloLens 2, and Magic Leap 2 have begun integrating iris authentication as a built-in form of user authentication. This authentication method is becoming more popular, especially when conventional modalities such as fingerprint or facial recognition are unavailable due to the device form factor. Moreover, conventional authentication methods, such as passwords and PINs, are susceptible to observation and reconstruction attacks within immersive environments [28].

The stability and uniqueness of irises make them a gold standard biometric and are well-suited for AR/VR environments [5]. However, the reliance on biometrics in AR/VR systems inherits existing risks associated with presentation attacks. Traditional presentation attacks have targeted iris recognition systems through physical means, such as a printed out iris image or a textured contact lens to mimic a real iris [42, 50]. More recently, digital attacks have emerged where adversaries remove the need for physical props by generating synthetic biometric data to deceive recognition systems [34].

This thesis examines the emerging threat of purely digital presentation attacks, driven by ad-

vances in generative modeling. Recent techniques, such as Generative Adversarial Networks (GANs) and diffusion models, have made it increasingly possible to synthesize realistic iris imagery, undermining the long-standing assumption that iris patterns are inherently difficult to replicate. Unlike traditional attacks that require physical artifacts, digital spoofs enable scalable, remote attacks against biometric systems. As AR/VR platforms increasingly rely on continuous biometric authentication, the ability to generate high-quality, adversarial iris images raises serious concerns about the long-term security and resilience of these technologies.

2.2 Adversary Capabilities and Assumptions

In this thesis, digital presentation attacks are defined as cases where an adversary has access to at least one high-quality image of a victim's iris. Unlike traditional physical attacks, where the adversary must use artifacts such as printed eyes or textured contact lenses, digital attacks use synthetic imagery to impersonate a user remotely. Although irises have been viewed as a highly secure and private identification method, recent developments challenge this confidence. There are several real-world situations where an adversary could obtain iris images without a user's consent. This is outlined below:

- **Social Media and Public Image Leakage**: High-resolution iris textures are often captured in selfies, group photos, and professional images shared on social media platforms. An adversary may apply super-resolution enhancement tools to extract usable iris data from these images. Alonso-Fernandez et al. [2] demonstrated that deep learning-based super-resolution models can reconstruct high-fidelity iris patterns from low-quality facial images. Ribeiro et al. [41] showed that upscaled iris images can reach sufficient quality for reliable biometric recognition.

- **Surveillance Systems**: Modern surveillance infrastructure includes long-range iris scanners capable of capturing detailed eye imagery from a distance and in non-cooperative scenarios. Lee and Park [30] showed that iris recognition remains feasible even under poor quality or long-range capture conditions.

- **Leakage from Generative Machine Learning Models**: When diffusion models or GANs are fine-tuned on biometric datasets, they can memorize and reproduce examples from their training data. Carlini et al. [9] revealed that diffusion models trained on sensitive data can regurgitate memorized identities. Our work showed that diffusion models trained on CASIA-Iris-Thousand [36] could regenerate synthetic images nearly indistinguishable from the original identities. These privacy risks are further evaluated in Chapter 4.

- **White-Box Access via AR/VR Headsets**: Modern AR/VR devices such as the Meta Quest Pro include internal, user-facing cameras used for eye-tracking and biometric interaction. In some configurations, such as when developer mode is enabled, these camera streams may be accessible to users or applications with elevated permissions. For example, the VRBiom dataset [27] was constructed by recording raw periocular video streams at 72 FPS using the Quest Pro's internal cameras. This suggests that in certain white-box threat scenarios, an adversary with developer access could extract high-resolution iris data directly from the device's internal sensor pipeline.

- **Hardware Embedding in VR Headsets**: If an adversary had physical access to a victim's VR headset, they could embed a hidden camera inside the headset without the user's knowledge. This kind of attack bypasses the need for software-level defenses entirely. This also allows for high-quality close-range images of the victim's iris. This is a serious concern in environments where VR headsets are shared and cannot be fully trusted.

Based on these factors, the adversary is assumed to:

- Possess a high-resolution iris image of target individuals through public image scraping, surveillance, or ML leakage.
- Apply image enhancement techniques to improve image quality.
- Utilize pre-trained generative models (e.g., diffusion, GAN) to generate spoofing samples.

2.3 Threat Implications

Successful digital presentation attacks against iris-based authentication systems can have serious consequences for both individuals and organizations. These attacks can grant adversaries unauthorized access to personal data, virtual workspaces, financial accounts, and other sensitive resources tightly integrated into the immersive computing experience. As AR/VR devices increasingly blur the boundary between personal and professional spaces, the security stakes of biometric spoofing grow even higher.

A successful spoof allows an adversary to impersonate a target user, gaining unauthorized access to sensitive communications, financial information, cloud-stored data, or private virtual environments. In consumer applications, this could expose email accounts, social media profiles, or banking services [22]. In enterprise or medical contexts, an attacker could access confidential business documents, patient records, or secure internal meetings hosted in virtual platforms.

Moreover, the unique nature of iris biometrics introduces a fundamental privacy concern: biometric traits are difficult, if not impossible, to revoke once compromised [21]. Unlike

passwords or security tokens, a compromised iris pattern cannot simply be ”reset,” placing individuals at long-term risk once their biometric template has been leaked or successfully spoofed.

Chapter 3

Background

3.1 AR/VR Biometrics

3.1.1 Limitations of Existing VR Authentication Methods

Conventional authentication systems often involve traditional methods, such as personal identification numbers (PINs), pattern-based inputs, and password entry. Nevertheless, such approaches raise several usability and security issues in VR. First, typing PINs or tracing gestures in VR often stops the user experience and presents them with a virtual input surface, interrupting immersion and workflow. Second, these interactions are frequently interpretable from gaze tracking or the patterns of shoulder/head movements, rendering them vulnerable to side-channel attacks [28]. Furthermore, it is difficult to revoke biometric traits once compromised [21]. Leaked iris data, unlike passwords, cannot simply be reset by users.

3.1.2 Related Work on Iris Biometrics in AR/VR

Recent years have seen the rise of integrating iris recognition in AR/VR headsets to enable secure authentication. Devices such as the Apple Vision Pro, Magic Leap 2, and Microsoft HoloLens 2 use iris-based authentication, rather than traditional passwords or PINs [4, 29, 46].

Nevertheless, academic studies on iris recognition still concentrate on the controlled datasets, such as the CASIA database [36], which do not reflect the lighting and gaze variability inherent in AR/VR environments. A limited number of efforts, such as VRBiom [27] and MagicEyes [52], have begun to capture eye-tracking and periocular data under AR/VR conditions using internal user-facing cameras. These datasets provide a foundation for evaluating iris recognition robustness in more realistic immersive settings.

3.2 Iris Spoofing

3.2.1 Physical Presentation Attacks

Biometric authentication systems have long struggled with presentation attacks, where an adversary tries to deceive the system by presenting fake biometric traits, such as printed iris images or textured contact lenses. Although iris recognition has been widely regarded as one of the most secure biometric modalities, thanks to the iris's stability and uniqueness, early research showed that it was not immune to spoofing. High-resolution printouts and artificial irises could bypass these systems, highlighting the need for liveness detection techniques to distinguish genuine users from fake samples.

Eye-Patch Presentation Attack and Liveness Detection

One of the first documented threats to iris recognition systems involved printing iris images, called iris print attacks. In this method, an adversary presents a high-resolution printout of a target's iris, often with a small aperture cut into the center to expose the pupil and mimic natural light reflections. This setup enables the spoof to bypass conventional iris recognition systems, which typically rely on Hamming distance metrics to assess similarity

between binary iris templates.

Hamming distance is a widely used metric in iris biometrics that quantifies the proportion of differing bits between two normalized iris codes. Recognition systems based on Daugman's algorithm assume that authentic iris samples exhibit low intra-class Hamming distances due to the inherent stability and uniqueness of iris textures [14]. However, print-based attacks have demonstrated that these assumptions can be violated. Specifically, the presentation of high-quality printouts can yield binary templates that closely resemble genuine samples, particularly when combined with techniques that simulate live pupil characteristics, such as light reflection.

These attacks were successful due to several factors: the use of high-resolution printed iris images capable of producing low Hamming distances during template matching; the inclusion of a pupil cutout that mimicked real light reflections, deceiving pupil detection mechanisms; and the absence of motion-based liveness detection, which allowed static samples to be treated as genuine biometric input.

To address this vulnerability, researchers explored alternative authentication mechanisms that incorporated dynamic features, such as eye movement signals, to verify *liveness* within a presented sample.

Eye Movement-Based Defenses

Rigas and Komogortsev introduced one of the first eye movement-driven liveness detection mechanisms to counteract print-based iris attacks [42]. Their approach leveraged that live human eyes exhibit unique gaze patterns, which a static printout cannot replicate. Their methodology:

- Analyzed gaze estimation distortions introduced by printouts during authentication.

- Identified inconsistencies in pupil motion between a real eye and a printed image.
- Achieved a maximum classification accuracy (ACR) of 96.5% and a minimum equal error rate (EER) of 3.4%, demonstrating its effectiveness in detecting spoofing attempts.

Raju et al. (2022) later built upon this foundation by integrating deep learning models, such as ResNet-based classifiers, to further enhance the detection of iris print attacks using eye movement signals [40]. Their method:

- Utilized deep learning to distinguish between natural eye movements and the lack of movement in spoofed irises.
- Required only 1.5 seconds of eye movement data for classification, improving efficiency in real-world deployment.
- Outperformed traditional statistical methods in terms of accuracy, making it a state-of-the-art approach for detecting iris presentation attacks.

Despite these advancements, presentation attack detection remains an unsolved problem. As noted by Czajka & Bowyer, *"One conclusion from this is that presentation attack detection for iris recognition systems is not yet a solved problem"* [13].

Other Physical Presentation Attacks

While iris print attacks were the first major form of biometric spoofing, additional physical presentation attacks have emerged over time, necessitating ongoing research into biometric security:

- **Textured Contact Lenses**: Attackers wear custom-designed contact lenses that

mimic the structure of a target's iris, fooling recognition systems. These lenses introduce artificial texture patterns that can deceive iris recognition algorithms [16].

- **Artificial Eyes**: Advanced prosthetics designed with realistic iris patterns have been tested against iris recognition systems. Research has shown that mechanical replicas can disrupt gaze estimation and liveness detection models, posing significant security challenges [26].

- **Replay Attacks**: Pre-recorded video sequences of a legitimate user's eye movements are used to bypass liveness detection. This type of attack is particularly effective against systems that rely solely on 2D imaging without verifying live interactions [39].

- **Mechanical Replicas**: Some studies have explored the feasibility of *robotic eye simulations* that attempt to mimic human gaze behavior. These sophisticated attacks introduce anomalies in gaze estimation algorithms, challenging traditional biometric defenses [25].

3.2.2 Digital Presentation Attacks

Iris recognition systems work off the premise that iris patterns are difficult to recreate by traditional means. Irises make a gold biometric standard due to their uniqueness and stability over time. Prior work within Presentation Attack Detection (PAD) has shown a high success rate in the context of detecting printed images, contact detection, or prosthetic eyes [44, 50]. However, prior work also shows that these detection mechanisms have a particular vulnerability against digital presentation attacks.

IrisSwap

Specifically, IrisSwap [34] revealed vulnerabilities within iris recognition systems by successfully using a segmentation-based approach, leveraging the rubber sheet model and a double U-Net architecture to extract and swap iris textures between images. This method successfully bypassed iris recognition along with a liveness detection model that classifies eye movements to verify human input. This attack was a significant advancement in biometric security research, showing that a purely digital attack could bypass authentication metrics without requiring *physical access* to a target's eye.

IrisSwap follows a segmentation-based pipeline that extracts and replaces iris textures using a rubber sheet model transformation. While this method successfully bypasses iris authentication, it also introduces visible artifacts where the swaps were visibly apparent.

Additionally, IrisSwap required substantial computational resources. The paper reports that due to its multi-step pipeline, including segmentation, iris normalization, and warping, the attack had high computational overhead, reducing real-time applicability. The pipeline's inference rate dropped as low as 3Hz in real-time deployment scenarios, making it impractical for live spoofing attacks [34].

These limitations highlight the need for a more adaptive, generative approach to produce high-quality, realistic swaps while maintaining efficiency and feasibility for real-world attacks. This leads to exploring GAN-based techniques, which learn from data distributions and generate realistic, seamlessly blended iris textures instead of relying on mechanical transformations.

Since IrisSwap demonstrated a fundamental vulnerability in digital iris presentation attacks, an important next step is to streamline the swap process by leveraging machine learning models, such as GANs, to improve the realism and efficiency of these attacks.

GAN-Based Iris Synthesis

While segmentation-based methods like IrisSwap successfully bypassed traditional iris authentication systems, they suffered from rigid transformations and segmentation artifacts, leading to unnatural swaps and high computational overhead. Even though these images were able to bypass iris authentication, it's clear to the human eye that a swap is occurring. These limitations prompted researchers to explore GAN-based approaches to generate more realistic iris-swapped images while preserving biometric identity.

Recent advancements in generative models have allowed models to transform and manipulate images within biometric attacks. GAN-based models have been widely used for face manipulation [31, 38], and similar techniques have been used for iris synthesis and spoofing. Some notable examples include:

DeformIrisNet: DeformIrisNet [24] employed a deep autoencoder architecture with an identity-preserving loss function to model iris texture deformations. While this method allowed for controlled deformations of the iris, it was not designed for realistic iris swapping, limiting its application in biometric attacks. Furthermore, it struggled with generating high-fidelity iris textures, often introducing blending artifacts and unnatural boundary mismatches.

iWarpGAN: iWarpGAN [53] aimed to separate identity from style, allowing attackers to modify certain iris characteristics while preserving biometric consistency. However, this method lacked fine-grained control over periocular features, leading to artifacts appearing at the iris boundary. This limitation made it more detectable by iris recognition systems that analyze iris-periocular consistency.

RaSGAN: RaSGAN [54] focused on synthetic iris image generation rather than direct iris swapping. While this model could generate high-quality fake irises, it was not optimized for

iris-swapping tasks, meaning that pupil misalignment and shading inconsistencies made the images more susceptible to Presentation Attack Detection (PAD) techniques.

GAN-Based Iris Recognition Approaches

Traditional iris recognition systems are built on the assumption that iris patterns are unique, stable over time, and difficult to reproduce. Early systems relied heavily on handcrafted feature extraction techniques. One of the foundational approaches, Daugman's rubber sheet model [14], normalized the iris region into a fixed-dimensional polar representation and compared binary iris codes using Hamming Distance. While effective under controlled conditions, this method struggled to accommodate pupil dilation, lighting variability, and gaze deviations—limitations that become more pronounced under adversarial spoofing attempts.

To address these challenges, researchers turned toward deep learning-based recognition systems. DeepIrisNet [18] introduced a convolutional neural network architecture that extracted robust features directly from raw iris images, significantly improving recognition performance across sensors and acquisition conditions. Statistical normalization techniques [51] and biomechanical models of iris deformation [48] further sought to enhance template matching under realistic noise and variability.

More recently, the use of GANs has expanded beyond synthesis to the recognition domain. Models such as LiveDet 2024 [55] leveraged generative models to create synthetic iris samples for training more resilient spoof detection systems. These approaches use GAN-generated data to augment training sets for Presentation Attack Detection (PAD) networks, enhancing the ability to distinguish live irises from digitally generated or replayed fakes.

Collectively, these advancements highlight a shift in iris recognition: from handcrafted feature comparisons toward deep, data-driven embedding spaces that are increasingly fortified

against modern digital spoofing attacks.

3.3 Attacks on ML Models

Recent work in machine learning security has revealed numerous vulnerabilities through which models can leak, expose, or be manipulated by training data. These attacks are broadly categorized into several classes:

- **Membership Inference Attacks**: Attackers determine whether a specific data point was included in a model's training set by analyzing its output behavior [45].
- **Model Inversion Attacks**: Attackers reconstruct sensitive features about the training data (e.g., partial facial features) by querying a model and observing outputs [17].
- **Data Poisoning Attacks**: Attackers introduce malicious examples into the training data to influence the model's behavior at test time [6].
- **Training Data Leakage from Generative Models**: Generative models, particularly GANs and diffusion models, have been shown to memorize and regurgitate sensitive samples from their training datasets [9].

The implications of these attacks are particularly severe within the context of biometric security. Unlike passwords, biometric traits such as iris patterns are immutable—once compromised, they cannot simply be changed. Leakage of biometric training data compromises user privacy indefinitely.

While much of the prior work has focused on facial datasets, this thesis extends the privacy leakage threat to the domain of iris recognition. In Chapter 4, we adapt the extraction

methodology proposed by Carlini et al. [9] to a diffusion model fine-tuned on the CASIA-Iris-Thousand dataset [36]. We show that the synthetic outputs retain high similarity to real training images and can be accepted by iris recognition systems, revealing the feasibility of identity leakage from model parameters in the iris domain.

This memorization capability opens new attack vectors in biometric systems, using generative models for data augmentation, user modeling, or synthetic training set expansion. In AR/VR environments where iris data is collected continuously, such vulnerabilities pose serious long-term privacy risks if synthetic modeling pipelines are adopted without careful security controls.

Chapter 4

Methodology

4.1 Iris Leakage Attacks

Before discussing the experiments, it is essential to understand how we adapt techniques from prior leakage studies (such as Carlini et al. [9]) to the iris biometric domain. We aim to determine whether diffusion models memorize and leak training iris images, posing privacy risks for authentication systems.

Thus, our methodology has four key steps: dataset selection, fine-tuning strategy, extraction attack design, and evaluation through an iris recognition system.

4.1.1 Dataset

CASIA-Iris-Thousand is a subset of CASIA-Iris-V4 [36] created by the Smart Iris Recognition group of the Chinese Academy of Sciences. CASIA-Iris-Thousand contains 20,000 images of irises from 1,000 subjects, all anonymized. Each image is a grayscale JPEG created from a dual-eye iris camera. Intra-class variation also exists due to the specular reflection caused by certain subjects wearing eyeglasses.

4.1.2 Fine-tuning

We begin by fine-tuning Stable Diffusion, as it is one of the most popular diffusion models researchers widely use. We preprocess CASIA-Iris-Thousand by only taking the first left iris for each subject, the rationale being that having fewer repeated images in the fine-tuning set will give our model the best chance to *not* memorize irises.

We fine-tune Stable Diffusion through LoRA dreambooth. Dreambooth [43] is a model created by Google Research that is widely accepted as the state-of-the-art way to personalize text-to-image models like Stable Diffusion. Dreambooth can perform successfully with as few as 3-5 images and is noted to maintain high fidelity while preserving small details of a subject. Chambon et al. [10] also noted that dreamboothing is currently the most effective method for adapting pre-trained vision-language models to medical imaging domains, demonstrating this with chest CT scans.

LoRA [20] is a technique developed by Microsoft Research that greatly reduces the GPU requirements and the size of the trained weights by decreasing the number of trainable parameters for the model. In addition to significantly speeding up inference, LoRA is a technique widely used by multi-modal researchers, so using this technique reflects a realistic way for researchers to create synthetic iris data.

The prompt for fine-tuning is "A photo of a klwmv human's left iris". We use klwmv as a random identifier so Stable Diffusion can dissociate its current knowledge of human irises from the ones we fine-tune it with.

4.1.3 Iris Extraction

Once Stable Diffusion is fine-tuned, we query the model with the fine-tuning prompt and random initial noise several times. This gives us a set of candidate synthetic irises. We generated 500 to follow Carlini's paper.

After generating the candidate set, we conduct membership inference. The intuition for this is that if we use the same prompt with different initial noise and generate similar irises, they are likely to be memorized. Expressed succinctly,

$$Gen(p; r_1) \approx_\sigma Gen(p; r_2) \tag{4.1}$$

where Gen is the generated output from the model, p is the prompt, r is the initial noise, and σ is an arbitrary distance threshold. As for the distance itself, we use the tiled ℓ_2 distance, which takes the maximum ℓ_2 distance over several patches of the images. This is done to avoid false positives created by factors such as a similar background.

Carlini et al. also choose to identify cliques of images instead of pairs to avoid false positives. A clique is a subset of vertices in a graph where every vertex is connected to another vertex.

We construct a graph over the candidate set to accomplish this, creating edges if the distance between any two images falls under σ. We manually tune σ to 35 to construct cliques of sufficiently large size.

When deciding which cliques to use, we always take the cliques of maximum size from the graph. This ensures that we are looking at the irises that are most likely to have been memorized. These irises are then tested against our evaluation, an iris authentication scheme.

4.1.4 GIRIST Recognition Software

For our recognition software, we utilized GIRIST, an iris recognition software provided by GRUSOFT [12]. Initially, we wanted to mimic Tinsley et al. by choosing a commercial recognition software such as VerifEye [47], however, they had a paywall-protected access to their software. This led us to GIRIST.

GIRIST does not have a unique foundational methodology. It mirrors the underlying pipeline of other established iris recognition frameworks. For example, a notable MATLAB implementation by Masek and Kovsi [33] remains to be used within the academic community. The relevance of this approach can be seen in different works such as exploring innovative privacy-preserving techniques within biometric systems by John et al. [23], and Chaudhary and Pelz. [11].

GIRIST operates by initially loading in all the iris images from the CASIA-Iris-Thousand dataset. During this process, all 1000 subjects are authenticated under GIRIST. Under the hood, GIRIST uses a localization method to identify where the iris is in the image. The program then segments the iris, which essentially flattens the iris into a rectangular image. After segmentation, the iris is then normalized/denoised and enhanced to capture the true intricacies of the iris. Subsequently, the program converts the image to binary vectors used during matching/recognition. During the matching phase, these binary vectors are XOR'd, and the resulting string gives us the Hamming Distance.

The software sets a hamming distance threshold according to the intra-class distribution (comparison of images within one participant) and the inter-class distribution (comparison of images between participants). Our CASIA-Iris-Thousand dataset had a threshold set to .37. If a comparison between images is below this threshold, it's deemed that the two images are from the same individual. This threshold is set by maximizing pairwise accuracy while

minimizing the equal error rate (EER) of the intra- and inter-class distributions.

4.1.5 Results

Testing with GIRIST: The final extracted irises were tested against GIRIST, an iris recognition system that:

- Segments and normalizes the iris.
- Encodes the iris into binary vectors for matching.
- Computes the Hamming Distance to determine if the iris matches a stored identity.

Table 4.1 summarizes the pass rates of our extracted synthetic irises in GIRIST. Each value in the first column represents the number of subjects within the CASIA-Iris-Thousand dataset on which we fine-tuned Stable Diffusion. Each value in the second column represents the ratio of synthetic irises that matched a real identity in the training set. The denominator in "x in y" corresponds to the largest extracted clique's size, representing the most likely memorized set of synthetic irises identified through our graph-based analysis. Figure 4.1 presents examples of synthetically generated irises for different training set sizes. Each iris shown corresponds to an image within the training set, successfully bypassing the authentication system.

Table 4.1: Pass rates for synthetic irises successfully matched to real identities by GIRIST. The first column indicates how many subjects we fine-tuned the diffusion model on. The denominator in "x in y" represents the size of the largest extracted clique from the graph-based memorization detection step.

# of subjects in Training Set	Pass Rate (x of y)
5	0 of 5
100	1 of 11
200	1 of 20
1000	1 of 9

(a) 100 subjects (b) 200 subjects (c) 1000 subjects

Figure 4.1: Synthetically generated irises using a diffusion model fine-tuned on different numbers of subjects. Each image passed the GIRIST authentication software for their respective subset of subjects.

Findings: Our results demonstrate that:

- Even with only 100 training subjects, a synthetic iris matched a real training iris.
- The pass rate increased with more training subjects, suggesting that larger datasets make biometric leakage more probable.
- Diffusion models trained on iris data can regenerate real identities, posing serious privacy concerns.

4.2 Iris Swapping via GAN-Based Identity Transfer

Having established the feasibility of iris retrieval in Chapter 4.1, we now describe our methodology for conducting a digital iris presentation attack. This work proposed a GAN-based iris-swapping framework, which employs a Cross-Attention GAN to generate realistic iris swaps. Our contributions focus on generalizing the method to AR/VR iris datasets, improving robustness under domain shift, and evaluating cross-domain spoofing success.

4.2.1 Iris Recognizers

Deep learning models have shown promising results in iris recognition [35]. To evaluate the identity preservation of iris-swapped images, we experiment with two iris recognition models that share the same architectural backbone but differ in training strategy and evaluation criteria. Both models use a modified ResNet-50 architecture pre-trained on ImageNet [15], with the first convolutional layer adapted for grayscale input.

In addition to serving as identity verifiers during evaluation, these models are integrated into our GAN training pipeline as feature extractors. Specifically, they provide identity embeddings that are used to compute identity loss during the optimization of our generator network.

Classification-Based Recognizer with Cosine Similarity

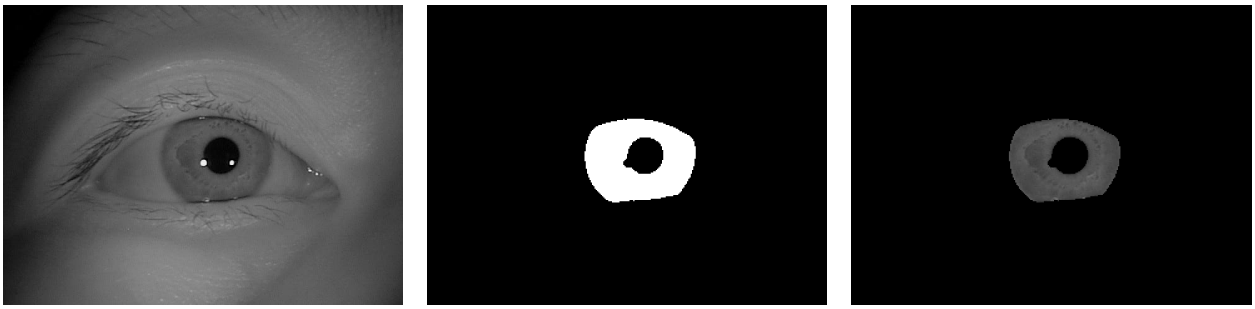
Prior to feature extraction, we apply a binary mask M to the input image I to isolate the iris region:

$$I_{\text{iris}} = I \odot M \tag{4.2}$$

where $\odot$ represents element-wise multiplication. This process is illustrated in Figure 4.2.

The iris recognition network follows a three-stage pipeline:

- *Feature Extraction*: We utilize ResNet-50's convolutional layers, omitting the classification head, to extract 2048-dimensional feature vectors.
- *Embedding Projection*: A fully connected layer projects feature vectors into a 128-dimensional space, followed by batch normalization and ReLU activation.



(a) Original eye image (b) Segmentation mask (c) Masked iris ($I_{\text{iris}} = I \odot M$)

Figure 4.2: Iris masking process used before feature extraction, shown on a sample from the CASIA-Iris-Thousand dataset [36]. The binary mask isolates the iris region in the input image, producing the masked iris image used by the recognizer.

- *L2 Normalization*: This ensures embeddings are uniformly distributed for similarity comparisons.

Given an input iris image I_{iris}, the network produces L2-normalized embeddings:

$$F_\theta(x) = \text{Norm}(f_{\text{emb}}(f_{\text{feat}}(I_{\text{iris}}))) \tag{4.3}$$

where: f_{feat} represents the ResNet-50 backbone, f_{emb} is the embedding projection layer, and Norm denotes L2 normalization. The embedding projection is defined as:

$$f_{\text{emb}}(z) = \text{ReLU}(\text{BN}(W_e z + b_e)) \tag{4.4}$$

where: $z \in \mathbb{R}^{2048}$ is the feature vector, and $W_e \in \mathbb{R}^{128\times2048}$ and $b_e \in \mathbb{R}^{128}$ are trainable parameters. During training, we incorporate an auxiliary classification head to map embeddings to identity labels:

$$\hat{y} = W_c f_{\text{emb}}(f_{\text{feat}}(I_{\text{iris}})) + b_c \tag{4.5}$$

where: $W_c \in \mathbb{R}^{C\times 128}$ and $b_c \in \mathbb{R}^C$ are classifier parameters, and C represents the number of identity classes within the dataset. The model is trained using a standard cross-entropy loss between the predicted logits $\hat{y}$ and the ground truth identity labels. This objective encourages the network to cluster embeddings from the same subject while maximizing separation between identities. The classification head is discarded during inference, and only the normalized embeddings are retained for biometric matching using cosine similarity.

During inference, iris matching is performed by computing the cosine similarity between normalized embedding vectors produced by two different iris images:

$$\text{Sim}(F_\theta(x_1), F_\theta(x_2)) = \frac{F_\theta(x_1) \cdot F_\theta(x_2)}{\|F_\theta(x_1)\|\|F_\theta(x_2)\|} \tag{4.6}$$

To establish a decision threshold for biometric matching, we compute cosine similarity scores between all genuine and impostor pairs within the validation set. We then determine the threshold that maximizes overall pairwise classification accuracy. This optimal threshold is used throughout our experiments to compute attack success rates (ASR). For completeness, we also report the Equal Error Rate (EER) and its associated threshold, reflecting the point at which the false acceptance and rejection rates are equal. Both thresholds are derived from the similarity score distributions and provide complementary insights into model performance.

Contrastive Learning-Based Iris Recognizer with EER-Driven Evaluation

To address the limitations of classification-based iris recognition in open-set or cross-domain settings, we employ a contrastive learning framework trained with the InfoNCE loss. This model is designed to generalize across subject identities and datasets by optimizing embedding similarity between matching pairs while pushing apart non-matching pairs. Unlike the

previous recognizer, this model does not rely on a classification head or explicit identity supervision, making it better suited for evaluating spoof attacks across AR/VR datasets such as OpenEDS2019 [19] and OpenEDS2020 [37].

As in the previous recognizer, we use a modified ResNet-50 backbone to extract 2048-dimensional feature vectors from masked iris images. These are then projected into a 128-dimensional embedding space via a fully connected layer with batch normalization and ReLU activation. Finally, the embeddings are L2-normalized:

$$F_\theta(x) = \text{Norm}(f_{\text{emb}}(f_{\text{feat}}(I_{\text{iris}}))) \tag{4.7}$$

The key difference lies in the training objective. Instead of optimizing a cross-entropy loss over identity labels, this model is trained using the normalized temperature-scaled cross-entropy (InfoNCE) loss [49], which operates on pairs of embeddings:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{\text{sim}(z_i, z_k)}{\tau}\right)} \tag{4.8}$$

where $\text{sim}(z_a, z_b)$ denotes cosine similarity between embeddings z_a and z_b, and τ is a temperature scaling factor that controls the softness of the distribution. For each anchor z_i, the loss encourages similarity with a corresponding positive sample z_j (i.e., another sample from the same subject), while pushing it away from all other samples z_k in the batch. The indicator function $\mathbb{1}_{[k \neq i]}$ ensures that the anchor is not compared with itself. This formulation enables the network to learn identity-preserving embeddings without requiring an explicit classification head.

To train the model, we use a batch sampling strategy that ensures multiple subjects are rep-

resented in each mini-batch, allowing for both positive and negative pair formation. Each image is first multiplied by its corresponding binary segmentation mask to isolate the iris region. The masked images are passed through the network to produce L2-normalized embeddings, and the InfoNCE loss is computed using subject identity labels to determine positive and negative relationships.

To enhance generalization across iris domains and increase robustness to visual variability, we apply a set of geometric augmentations during training. Each image undergoes random rotation, horizontal flipping, and random zoom/cropping, ensuring the learned embedding space remains stable under minor spatial transformations. This encourages the model to focus on identity-specific iris features rather than image layout or gaze angle.

We also use the MPerClassSampler to construct mini-batches that contain multiple samples from multiple subjects, enabling the loss to form both positive pairs (same identity) and negative pairs (different identities) within each batch. This sampling strategy is critical for effective contrastive learning and ensures that the InfoNCE loss operates on a meaningful distribution of similarities.

We monitor validation loss during training and compute the Equal Error Rate (EER) on the embeddings extracted from the validation set. We save the model checkpoint with the highest EER-based validation accuracy across epochs.

4.2.2 Generator Architecture Design

Our generator architecture takes the following inputs to perform precise iris swapping: the masked attacker eye image I_{eyeball}, the attacker's binary segmentation mask M_{attacker}, and the target victim iris image I_{iris}.

Our generator follows a modified U-Net architecture with three key innovations: the elim-

ination of skip connections, a dual-encoder structure, and a cross-attention module at the bottleneck.

Elimination of Skip Connections: By eliminating skip connections within our decoder path, our network achieves more effective feature transfer from the victim's iris. Removing traditional U-Net skip connections, we prevent residual features of the attacker's original iris from being retained within the generated iris.

Dual-Encoder Design: Inspired by [7], which used dual encoders for improved image reconstruction, we use a secondary encoder dedicated to processing the victim's iris image. This ensures that detailed iris features are effectively transferred while maintaining the anatomical consistency of the attacker's eye region.

Cross-Attention Module: At the bottleneck of the generator, we introduce a cross-attention module designed to fuse identity features from the victim's iris with the spatial context of the attacker's eye. This module uses the attacker's encoder output as the query, and the victim's iris features as both key and value. Multi-head dot-product attention is used to align and selectively transfer relevant iris textures. A residual connection adds the attention output back to the original attacker features:

$$F_{\text{fused}} = \text{Attn}(F_{\text{eyeball+iris}}, F_{\text{iris}}, F_{\text{iris}}) + F_{\text{eyeball+iris}} \tag{4.9}$$

This structure allows the model to dynamically incorporate fine-grained identity features while preserving the overall anatomical consistency of the eye region. Rather than simply copying texture, the attention mechanism facilitates precise spatial alignment between the victim's iris and the attacker's eye geometry, enabling realistic and seamless iris swaps.

Cross-Attention Module: At the bottleneck of the generator, we introduce a cross-

attention module designed to fuse identity features from the victim's iris with the spatial context of the attacker's eye. This module uses the attacker's encoder output as the query, and the victim's iris features as both key and value. Multi-head dot-product attention is used to align and selectively transfer relevant iris textures. A residual connection adds the attention output back to the original attacker features:

$$F_{\text{fused}} = \text{Attn}(F_{\text{eyeball+iris}}, F_{\text{iris}}, F_{\text{iris}}) + F_{\text{eyeball+iris}} \tag{4.10}$$

This structure allows the model to dynamically incorporate fine-grained identity features while preserving the overall anatomical consistency of the eye region. Rather than simply copying texture, the attention mechanism facilitates precise spatial alignment between the victim's iris and the attacker's eye geometry, enabling realistic and seamless iris swaps.

The decoder receives the fused feature map from the cross-attention module and reconstructs the full-resolution eye image through transposed convolutional layers. The output of the decoder is not directly used as the final output. Instead, it is blended into the original eye image using the attacker's binary iris mask to ensure only the iris region is replaced:

The final output is blended using a binary mask to ensure only the generated iris region is modified:

$$I_{\text{final}} = I_{\text{eyeball}} \cdot (1 - M_{\text{attacker}}) + I_{\text{gen}} \cdot M_{\text{attacker}} \tag{4.11}$$

4.2.3 Training Objectives and Loss Functions

The generator is optimized using a combination of adversarial and identity-preserving objectives to ensure that the synthesized eye images are both photorealistic and biometrically

consistent with the target identity.

Specifically, the generator is trained using two adversarial losses and two identity losses:

Adversarial Losses: We adopt a least-squares GAN (LSGAN) formulation to stabilize training and promote high-quality image synthesis. Two PatchGAN discriminators are employed:

- A *full-image discriminator* D that distinguishes between real and generated full eye images.
- An *iris-region discriminator* D_{iris} that focuses exclusively on the masked iris regions.

The adversarial loss for each discriminator is defined as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{I_r}[(D(I_r) - 1)^2] + \mathbb{E}_{I_g}[D(I_g)^2] \tag{4.12}$$

$$\mathcal{L}_{\text{iris}} = \mathbb{E}_{I_r}[(D_{\text{iris}}(I_r) - 1)^2] + \mathbb{E}_{I_g}[D_{\text{iris}}(I_g)^2] \tag{4.13}$$

where I_r and I_g represent real and generated images or iris regions, respectively.

Identity Preservation Losses: To encourage the generated iris to preserve the target identity, we employ a pretrained iris recognizer network. Let F_θ denote the pretrained feature extractor that maps masked iris images to L2-normalized embeddings. Two identity loss terms are computed:

- A cosine similarity loss between the feature embeddings of the generated and target iris:

$$\mathcal{L}_{\text{iris\_features}} = 1 - \cos(F_\theta(I_{\text{gen}}), F_\theta(I_{\text{iris}})) \tag{4.14}$$

- An L1 loss between the feature embeddings of the generated and target iris:

$$\mathcal{L}_{\text{iris\_f1}} = \|F_\theta(I_{\text{gen}}) - F_\theta(I_{\text{iris}})\|_1 \tag{4.15}$$

Total Generator Loss:

The total generator loss is computed by equally summing the four loss components, without applying explicit weighting factors:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{iris}} + \mathcal{L}_{\text{iris\_features}} + \mathcal{L}_{\text{iris\_f1}} \tag{4.16}$$

This composite objective jointly encourages the generation of realistic full-eye images, realistic iris regions, and strong biometric identity preservation in feature space.

4.2.4 Training Setup

Our attack pipeline is optimized using the Adam optimizer with a learning rate of 0.0002 and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All models are implemented in PyTorch and are trained for 200 epochs with a batch size of 32. We resize all images to 512x512 before passing them into the model to test the generalization of our methodology to other datasets of varying resolutions. Training was conducted on an NVIDIA A100-SXM4-80GB GPU.

4.2.5 Evaluation Metrics

To evaluate the effectiveness of our iris-swapping attacks, we compute the **Attack Success Rate (ASR)** at two different levels:

- **Sample-Level ASR:** The proportion of individual spoofed samples that exceed a biometric similarity threshold when compared against the victim's real iris embedding. This measures the overall strength of the impersonation on a per-image basis.

- **Subject-Level ASR:** For each attacker–victim pair, we compute the percentage of spoofed samples that are successful. If more than 50% of samples for a given victim exceed the similarity threshold, the attack is considered successful for that victim. The subject-level ASR is the proportion of victims for which this majority-vote spoofing attempt succeeds.

We adopt a leave-one-subject-out evaluation protocol. Each subject in the dataset is treated as an attacker, attempting to spoof the identity of every other subject using the iris-swapping generator. Neither attacker nor victim subjects are seen during the training of the generator or the recognizer. This setup reflects a realistic black-box biometric attack scenario on *unseen individuals*. This setup reflects a realistic black-box biometric attack scenario on *unseen individuals*.

Cosine similarity between deep iris embeddings is used to determine match success. The decision threshold is selected from a validation set to optimize either Equal Error Rate (EER) or pairwise classification accuracy. A spoofed image is successful if the similarity score between the generated iris and the target victim's iris exceeds this similarity threshold.

4.3 Datasets

4.3.1 CASIA

The CASIA-Iris-Thousand dataset [36] is a widely used benchmark in iris recognition research, developed by the Chinese Academy of Sciences. It consists of 20,000 near-infrared iris images collected from 1,000 subjects, equating to 10 images per eye per person. Each subject has images from both eyes captured across multiple sessions both with and without glasses.

Images in CASIA are captured under controlled indoor lighting conditions using near-infrared (NIR) cameras, resulting in high-contrast, texture-rich iris patterns with minimal environmental variability. Each image is grayscale and originally sized at 640x400 pixels.

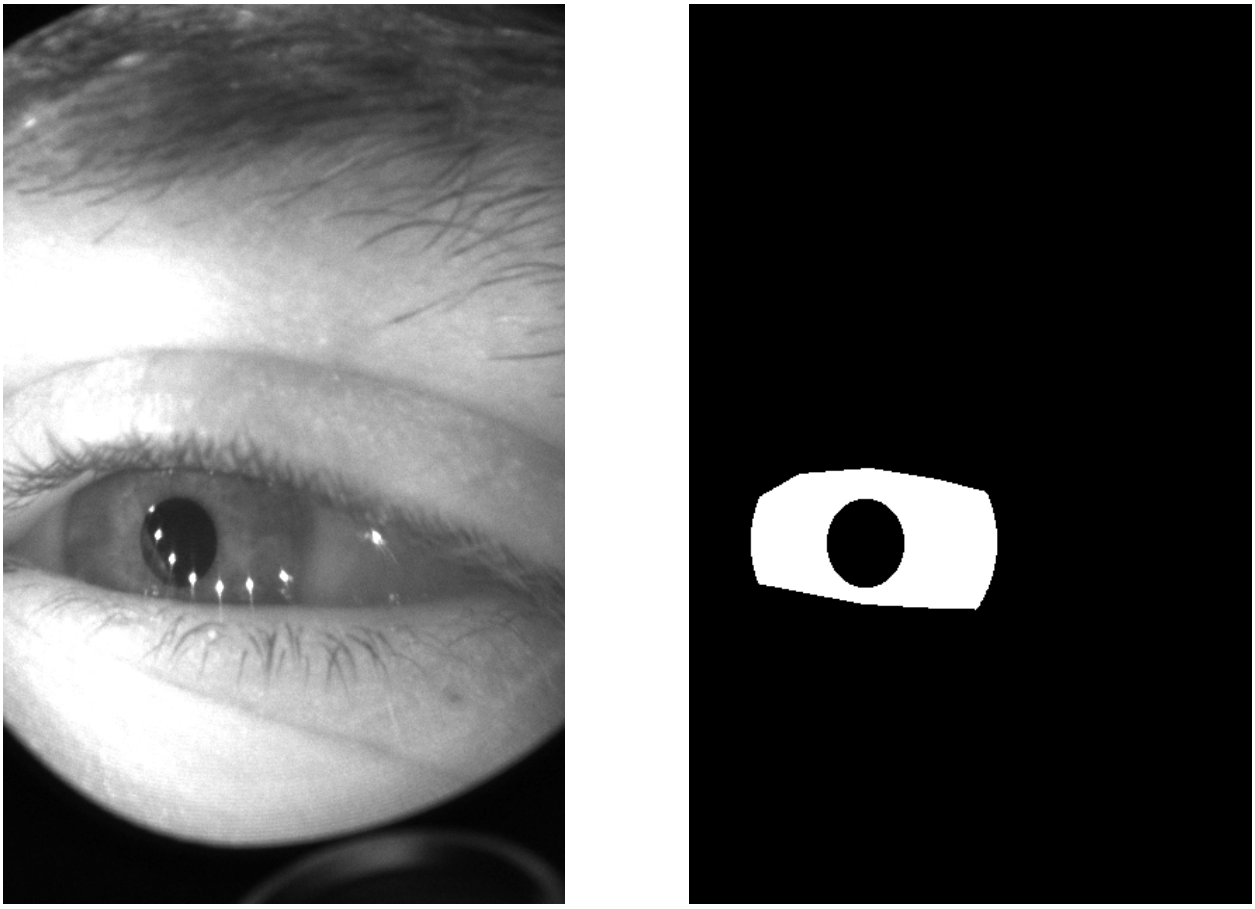
To generate iris segmentation masks for CASIA-Iris-Thousand, we used the open-source iris segmentation toolkit provided by the University of Notre Dame [1]. This toolkit outputs binary masks identifying the iris region, enabling us to construct precise (image, mask) pairs required for training and evaluation. An example of a mask & image within the CASIA dataset can be seen in Figure 4.2.

4.3.2 OpenEDS 2019

To train our iris-swapping models on VR-specific data, we use the OpenEDS2019 dataset [19], a large-scale dataset designed to advance eye-tracking applications in virtual reality (VR) settings. OpenEDS2019 consists of high-resolution near-infrared (NIR) eye images captured with head-mounted displays (HMDs). The dataset contains 356,649 images from 152 subjects, providing pixel-wise annotated segmentation masks for eye regions, including sclera, iris, and pupil.

The images in OpenEDS2019 have a resolution of 400x640 pixels and were recorded at 200 FPS. Each image is accompanied by precise semantic segmentation labels, making it particularly useful for iris recognition, gaze tracking, and biometric authentication tasks.

A major advantage of OpenEDS2019 over traditional iris datasets such as CASIA-Iris-Thousands is its realistic VR capture environment, which includes variable gaze directions, occlusions from eyelashes, and partial eye closures. These factors make it a more challenging and representative dataset for biometric authentication in head-mounted displays. An example of an image an mask pair is shown in Figure 4.3.



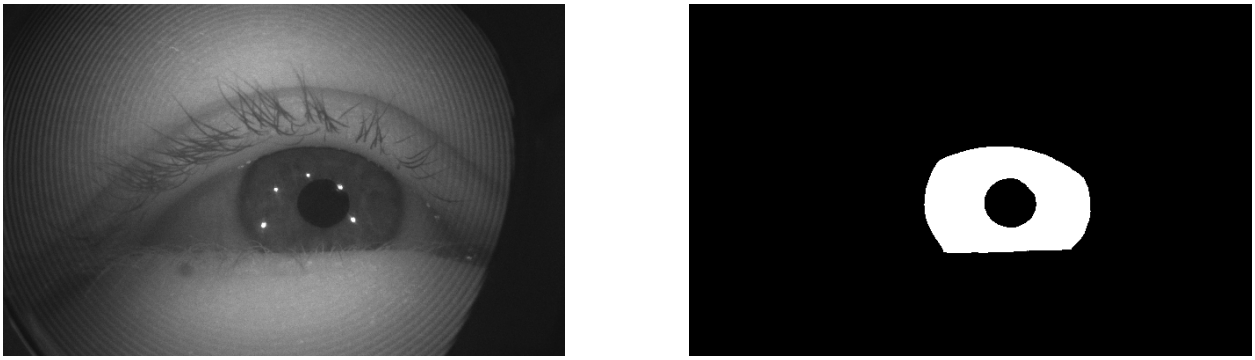
(a) Original eye image (b) Segmentation mask

Figure 4.3: Example image and segmentation mask pair from the OpenEDS 2019 dataset [19].

4.3.3 OpenEDS 2020

Facebook Reality Labs released the OpenEDS 2020 dataset [37] to support research in semantic segmentation of eye regions for AR/VR applications. It contains over 12,000 high-resolution 640x400 grayscale images of human eyes, captured using head-mounted devices equipped with infrared cameras. We use this dataset to evaluate the generalization of our attack pipeline.

Each image is accompanied by pixel-wise annotations for the iris, pupil, and sclera, making the dataset particularly well-suited for training and evaluating segmentation models in unconstrained gaze and pose conditions. The data features significant variability in head pose, gaze direction, eyelid occlusion, and lighting, providing a challenging testbed for biometric generalization. An example of this mask and image pair is seen in Figure 4.4.



(a) Original eye image (b) Segmentation mask

Figure 4.4: Example image and segmentation mask pair from the OpenEDS 2020 dataset [37].

4.3.4 Neon Dataset

To evaluate the generalizability of our iris-swapping model to VR eye images beyond the OpenEDS datasets, we conducted a user study using the Pupil Labs Neon eye tracker embedded within a Meta Quest 3 headset. This study was approved by the university's

Institutional Review Board (IRB) and involved collecting infrared eye-tracking data from participants performing various gaze-related tasks within a custom-designed Unity virtual environment. An example image and mask pair is shown in Figure 4.6.

Experimental Design

Our data collection was influenced by prior work in GazeBaseVR [32], a large-scale dataset initially developed for studying eye movement biometrics in VR. Eye movement behavior has been explored as a behavioral biometric modality and proposed as a countermeasure to spoofing attacks in iris-based systems. However, GazeBaseVR only provides gaze coordinates without corresponding eye video streams. Since our research focuses on image-based spoofing and requires high-frame-rate iris imagery, we developed custom Unity environments to capture both eye video and gaze data simultaneously.

Each participant performed a series of gaze-based tasks while the Pupil Labs Neon recorded grayscale eye videos (192×192 resolution) alongside synchronized gaze data and head pose information. This dataset enables analysis of biometric authentication, gaze tracking, and iris recognition under various viewing conditions.

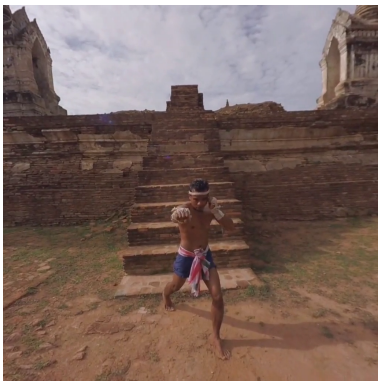
Participant Size and Demographics

A total of 30 participants (after excluding any incomplete or faulty recordings) were recruited for this study. Participants ranged in age from 18 to 35 years (M = 23.9, SD = 3.8). The study was conducted with a cohort of university-affiliated participants, including faculty and students. All participants had normal or corrected-to-normal vision. The gender distribution was as follows: Male (N = 16, 53.3%), Female (N = 13, 43.3%), and Non-Binary (N = 1, 3.3%).

Task Design

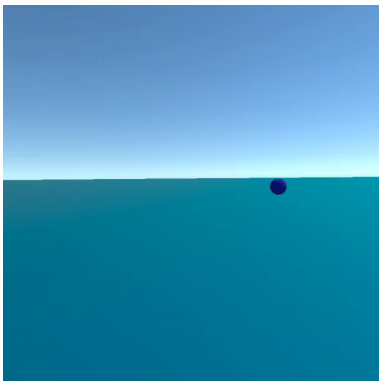
Participants completed six tasks, each designed to elicit distinct eye movement patterns (see Figure 4.5 for sample images from each task):

- **Video-Viewing Task:** Participants viewed four 30-second 360-degree videos and were instructed to explore the virtual environment naturally by freely shifting their gaze.
- **Image-Viewing Task:** Four 360-degree images were displayed for 30 seconds each, allowing participants to look around the scene freely.
- **Smooth Pursuit Task:** A floating sphere moved smoothly across the user's field of view at a constant velocity. Participants were instructed to track the sphere with their eyes while keeping their head still. This task lasted 30 seconds.
- **Random Saccade (RAN) Task:** A sphere moved to random locations within the user's field of view with varying velocities and dwell times. Participants were asked to follow the sphere each time it changed position quickly. This task lasted 30 seconds.
- **Vergence Task:** A sphere moved back and forth along the depth axis to induce vergence movements, where both eyes converge or diverge to maintain focus. This task lasted for 30 seconds.
- **Reading Task:** Participants read three randomly generated text passages at their own pace. Upon finishing a passage, they pressed a button to proceed to the next one.

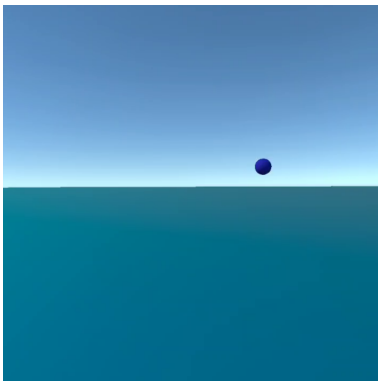


(a) Video-Viewing

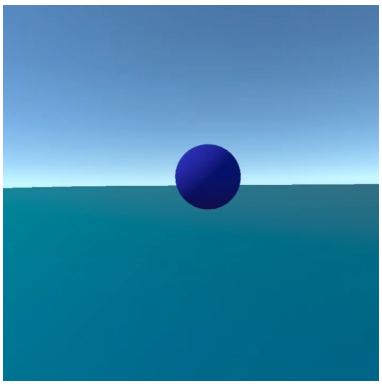
(b) Image-Viewing



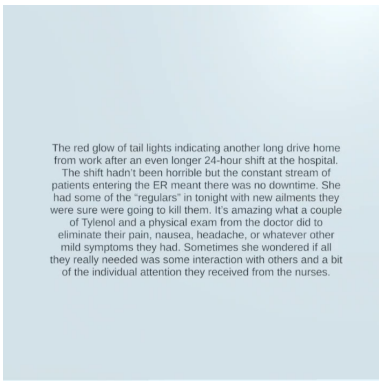
(c) Smooth Pursuit



(d) Random Saccade



(e) Vergence



(f) Reading

Figure 4.5: Sample screenshots from each of the six VR gaze tasks performed during data collection.

(a) Original eye image

(b) Manually labeled segmentation mask

Figure 4.6: Example image and segmentation mask pair from the collected Neon dataset.

Chapter 5

Results

This chapter evaluates the effectiveness of our proposed iris-swapping attack pipeline across several biometric datasets.

We initially explored evaluation on the Neon dataset, which captures realistic VR-specific eye imagery using the Pupil Labs Neon embedded within a Meta Quest 3 headset. However, the low native resolution (192×192 pixels), segmentation difficulties, and domain mismatch prevented effective application of our swapping pipeline. As a result, Neon is excluded from our quantitative evaluations, and we focus the remainder of this chapter on datasets with higher resolution and more reliable segmentation: CASIA-Iris-Thousand, OpenEDS2019, and OpenEDS2020.

We first establish a baseline using the CASIA-Iris-Thousand dataset, a widely adopted benchmark in traditional iris recognition. Since CASIA images are captured in controlled near-infrared conditions, this experiment validates our architecture's core components under ideal circumstances.

Next, we isolate our evaluation to a single VR-specific dataset, OpenEDS2019, which captures eye images in a head-mounted display environment. Unlike CASIA, the OpenEDS dataset includes significant variation in gaze angles, eyelid occlusion, and head pose—conditions that better reflect real-world biometric capture in AR/VR systems. In this experiment, we train the iris recognizer and the generator solely on OpenEDS data and

test the attack pipeline in-domain on unseen subjects.

Lastly, we evaluate the generalization capacity of our iris-swapping pipeline by training the model on a combined dataset consisting of CASIA and OpenEDS2019. We first assess attack success within each domain separately to verify performance under mixed-domain training. Then, we test the model's ability to generalize to a completely unseen dataset, OpenEDS2020, which introduces a new subject pool, different resolutions, and distinct lighting conditions. This evaluation enables us to measure whether our generator and recognizer, trained only on known domains, can produce convincing identity swaps that pass recognition thresholds on data captured from a VR sensor on which they were not trained.

5.1 Neon

While the Neon dataset captures realistic VR-specific eye images using the Pupil Labs Neon eye tracker embedded within the Meta Quest 3 headset, it introduced several significant challenges for our attack methodology. First, the native image resolution is limited to 192×192 pixels, much lower than datasets like CASIA or OpenEDS. Standard segmentation models struggled to accurately identify the iris region, likely due to the low pixel count and domain mismatch. These models are typically trained and benchmarked on publicly available datasets such as CASIA and OpenEDS, which feature higher-resolution, forward-facing images. In contrast, the Neon device's unique sensor placement at the bridge of the user's nose introduces lighting artifacts and gaze angles not seen during segmentation model training. Furthermore, our similarity models trained on other datasets failed to generalize effectively to Neon images. Together, these challenges led us to exclude the Neon dataset from quantitative spoofing evaluations in this thesis.

To illustrate this challenge, Figure 5.1 shows the distribution of cosine similarity scores

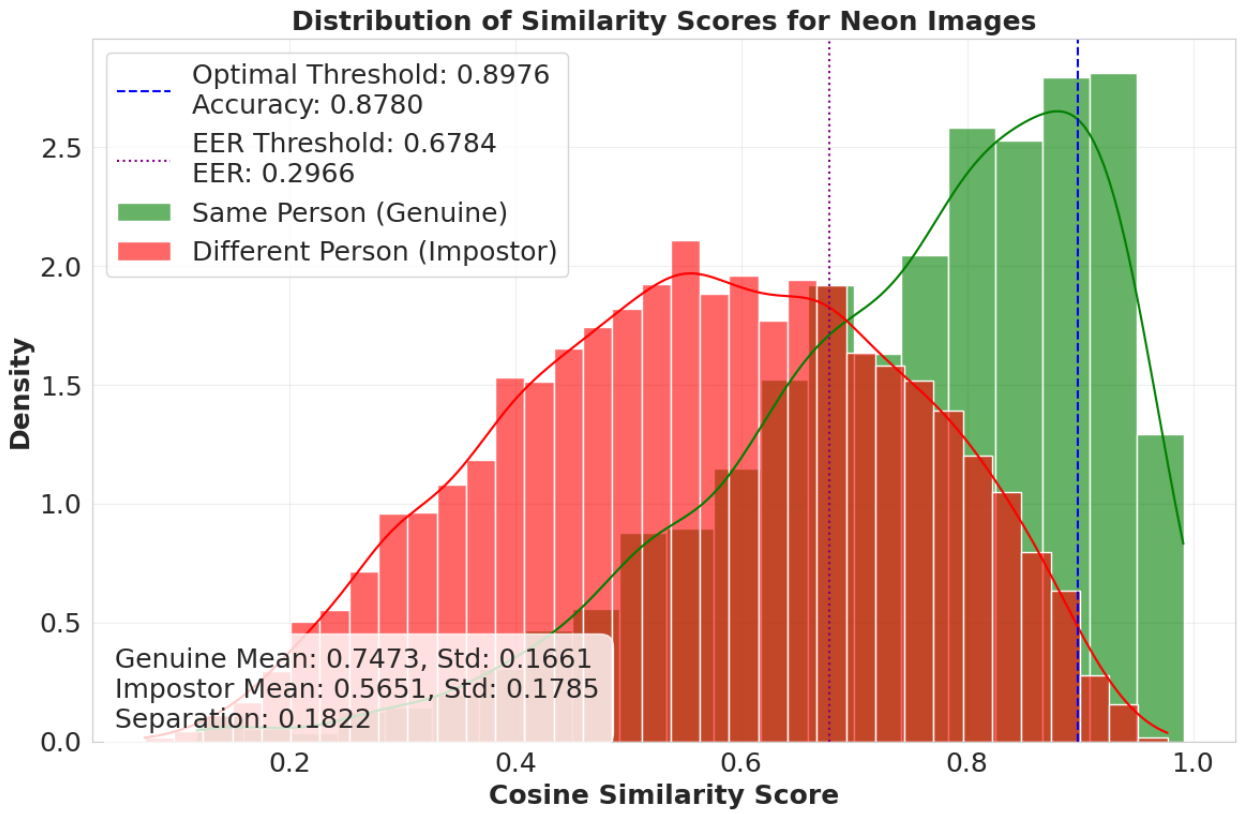


Figure 5.1: Distribution of cosine similarity scores between genuine and impostor pairs in the Neon dataset using the contrastive recognizer trained on CASIA and OpenEDS2019. The high overlap in distributions (29.66% EER) indicates that our contrastive-learning recognizer does not generalize well to Neon data.

between genuine and imposter pairs in the Neon dataset using our contrastive recognizer. To ensure proper iris segmentation, we manually labeled 10 images per subject for 300 iris segmentation masks for this comparison. Compared to the other datasets, there is a substantially greater overlap between the two distributions, indicating reduced identity separability on the neon dataset.

One promising direction for future work would be applying domain-specific super-resolution techniques aimed at enhancing iris detail without introducing artifacts that would compromise identity recognition [3]. However, incorporating such techniques was beyond the scope of this thesis and remains an open area of exploration.

5.2 CASIA

The CASIA-Iris-Thousand dataset serves as the foundational testbed for our evaluation. Due to its controlled imaging conditions, high-resolution near-infrared captures, and wide subject diversity, CASIA provides an ideal environment to benchmark the core components of our pipeline. In this section, we assess both the identity recognition performance of our classification-based iris recognizer and the effectiveness of our iris-swapping attacks when trained and tested exclusively within the CASIA domain.

5.2.1 Identity Recognition Performance

Classification-based Recognizer

We begin our evaluation by assessing the effectiveness of the Classification-Based Iris Recognizer trained solely on the CASIA-Iris-Thousand dataset [36]. Each eye in the dataset (left and right) is treated as a separate identity label during training, meaning that the model learns to distinguish between even the two eyes of the same individual. Figure 5.2 shows the cosine similarity distributions of same-subject (genuine) and different-subject (imposter) iris image pairs within unseen subjects within the CASIA dataset. The model achieves an equal error rate (EER) of 6.08% with a model accuracy of 95.43%. We set the optimal threshold be 0.6681 based on pairwise classification accuracy. These results indicate that the classification-based recognizer effectively distinguishes between real identities and provides a reliable reference point for evaluating spoofed images generated by our swap model.

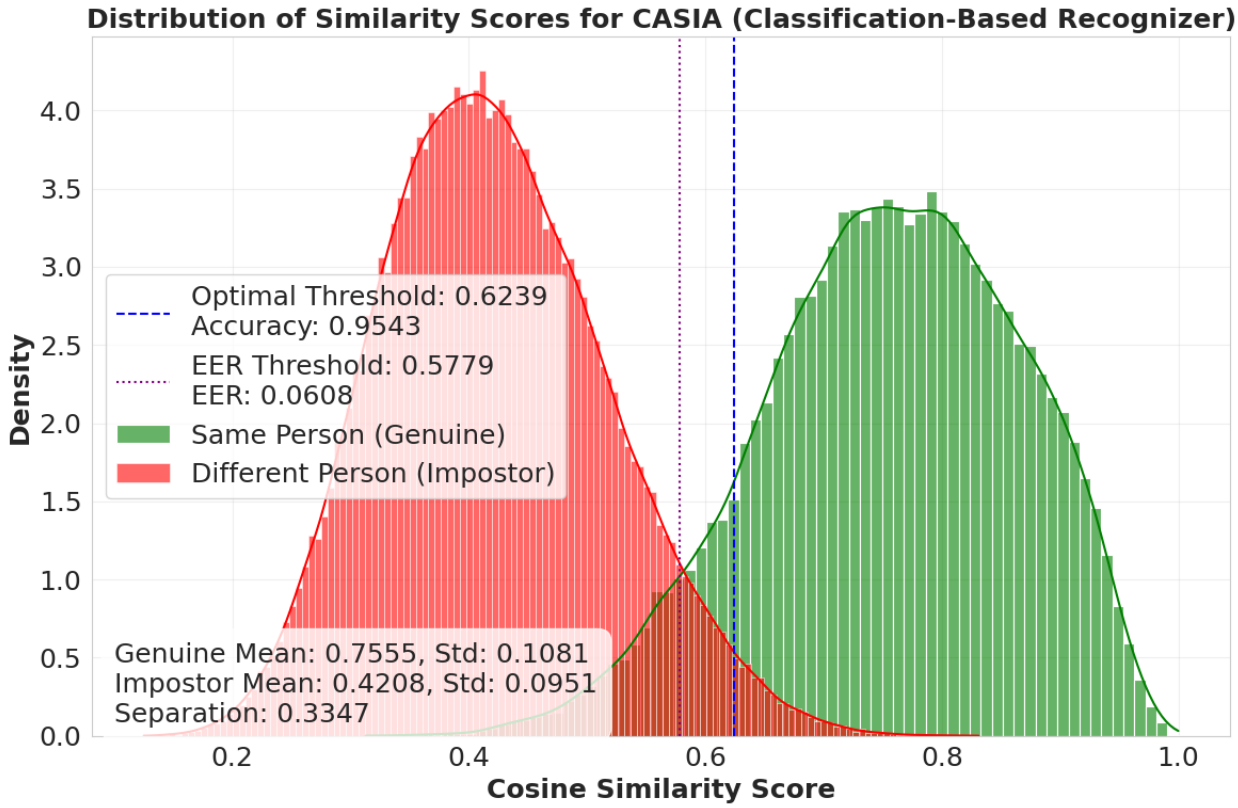


Figure 5.2: Distribution of cosine similarity scores between genuine and impostor pairs on unseen individuals in the CASIA dataset using the classification-based recognizer. We sample up to 30 same-class and different-class comparisons per image to ensure a representative evaluation for each unseen individual. The low EER and high classification accuracy demonstrate that the model is highly effective at distinguishing between matching irises (same identity) and non-matching irises (different identities).

Hamming Distance Comparisons

In addition to our cosine-based recognizer, we also evaluate identity using the HDBIF matcher [1], which operates on binary codes and Hamming distance comparisons. Unlike the classification-based recognizer used during training, this system encodes iris regions into binary codes and compares them using Hamming distance. Hamming distance remains one of the most widely used metrics in traditional iris recognition systems due to its simplicity, efficiency, and strong empirical performance in controlled environments. It's a foundational way to evaluate attack pipelines due to its simplicity, efficiency, and evaluation performance. CASIA suits this metric due to its frontal-view images and no gaze variation.

To bypass recognition under Hamming distance, we must first establish a threshold that separates genuine matches from impostor attempts. This involves calculating the distribution of Hamming distances across the entire CASIA dataset. As shown in Figure 5.3, the HDBIF matcher achieves a strong separation between these two distributions, with an equal error rate of 3.98% and a classification accuracy of 96.33%. We use a decision threshold of 0.38, which maximizes pairwise classification accuracy on genuine vs. imposter distributions. The low EER and clear margin between classes prove that Hamming Distance is still a strong metric in ideal datasets such as CASIA.

Contrastive Recognizer Limitations

We also experimented with a contrastive recognizer using InfoNCE. However, the limited size and narrow domain of CASIA led to poor generalization with a higher overlap between the imposter and genuine distributions. Therefore, we did not continue using this recognizer in further CASIA-specific evaluation.

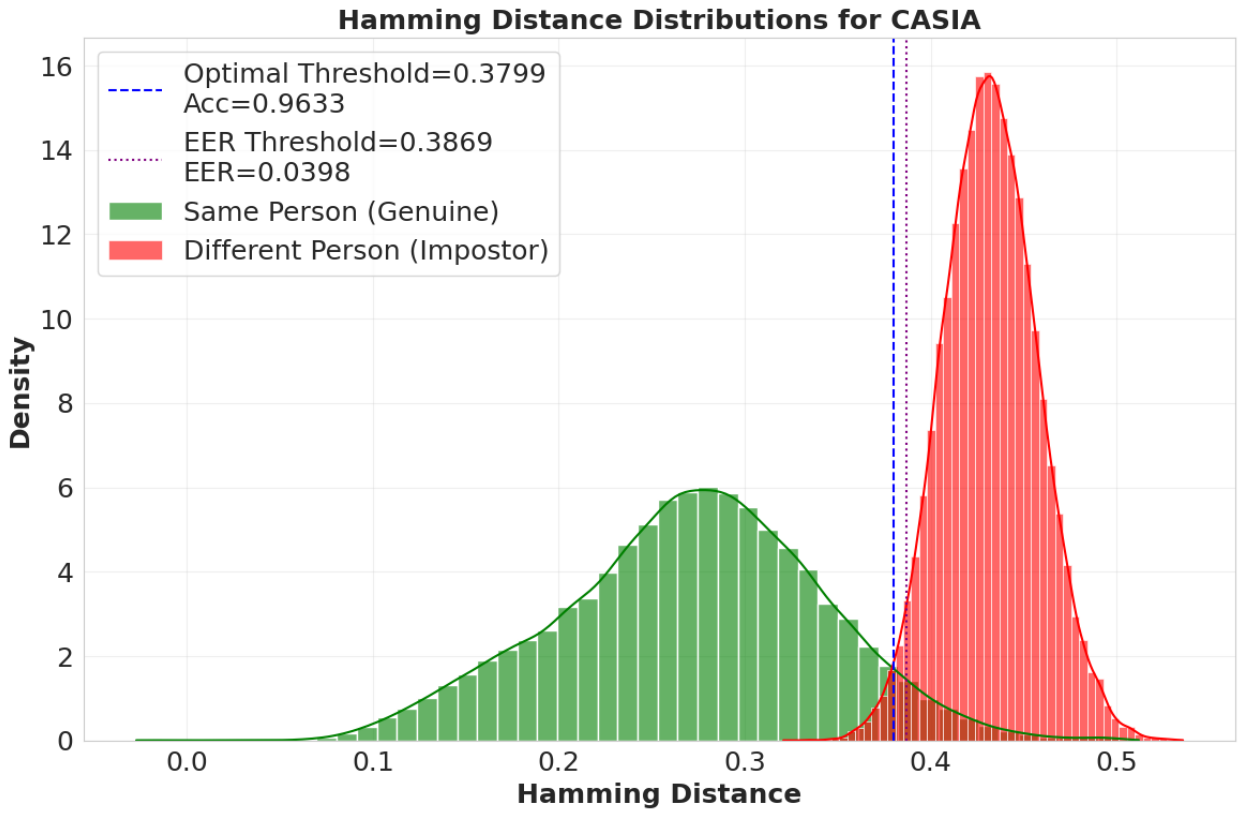


Figure 5.3: Distribution of Hamming distances between genuine and impostor pairs in the CASIA dataset using the HDBIF matcher.

5.2.2 Swap-Based Attack Evaluation

To evaluate the effectiveness of our digital presentation attack, we train a generator based on the cross-attention GAN architecture using the filtered CASIA-Iris-Thousand dataset. For each training sample, we construct an attacker input by masking out the iris region from their eye image, producing an inpainted eyeball region. The generator takes three inputs: (1) the masked attacker image (eyeball only), (2) the attacker's binary iris mask, and (3) the isolated iris region of a different subject (the victim). The generator is trained to synthesize a new eye image that realistically blends the victim's iris into the attacker's eye context while preserving anatomical structure.

We evaluate attack effectiveness under two different settings:

- **White-box attack:** The generator is evaluated against the same classification-based iris recognizer used during training. Since the generator receives feedback from this model through identity-based loss terms, it represents a best-case scenario for spoofing success.

- **Black-box attack:** The generator's outputs are evaluated against a traditional iris recognition system that compares binary iris codes using Hamming distance. This recognizer is not used during training, and the generator has no access to its gradients or internal structure.

These complementary evaluations allow us to assess both the targeted attack capability of our model and its generalizability to unseen recognition pipelines.

Training Setup

Before training, we apply a two-pass filtering process to the CASIA dataset to ensure label balance and mask quality. In the first pass, we compute the median iris area for each subject-eye combination (e.g., left eye of subject 800) using the binary masks. In the second pass, we discard samples whose iris area falls below that subject-eye's median. We further restrict training to classes with at least 5 valid samples to ensure sufficient intra-class diversity. This process removes noisy or poorly segmented examples.

Evaluation Setup

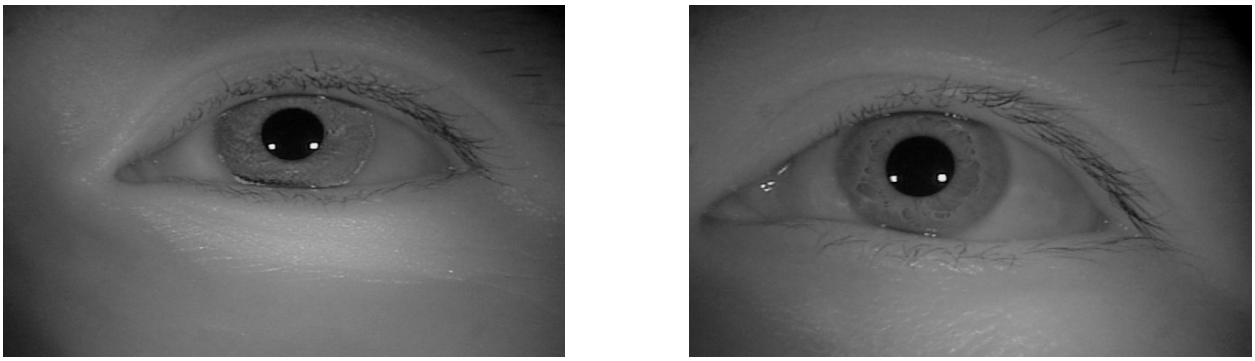
To evaluate the attack success rate (ASR) against our two conditions using, we generate spoofed images using our trained generator. For data filtration, we follow the same protocol as our swap model training setup, where we discard poorly segmented iris images. To maximize the quality within our swaps, our victim irises are selected by choosing the largest segmented iris area for that victim. Specifically, for each victim, we compute the pixel count within the binary segmentation mask and select the image where the iris region has the largest area. This follows our assumption that the adversary already possesses a high-quality image of that victim's iris.

Formally, given a set of victim samples $\{(I_k, M_k)\}$, we compute the iris area as $\text{Area}(M_k) = \sum M_k$, and select:

$$(I_{\text{victim}}, M_{\text{victim}}) = \arg\max_k \text{Area}(M_k) \tag{5.1}$$

This evaluation setup ensures that each subject serves as an adversary for every other individual, resulting in a comprehensive and balanced measurement of Attack Success Rate

(ASR). Importantly, all test-time subjects are unseen during the training of both the generator and the recognition models. All image pairs within our evaluation set were set aside during the training of our similarity model and generator. An example of this is seen within Figure 5.4.



(a) Spoofed image (attacker)

(b) Victim iris used in swap

Figure 5.4: Example spoofed image generated by blending the attacker's periocular region with the iris from a different subject. This victim's iris was selected based on having the largest segmented area.

White-Box ASR (Classification-Based Recognizer)

To evaluate the effectiveness of our attack within a white-box environment, we test the generator against the same classification-based iris recognizer during training. Following the same evaluation protocol described earlier, each of the 197 unseen subjects from the filtered CASIA test set acts as an attacker. For each attacker, spoofed eye images are generated by combining their periocular region with a high-quality iris sample from every other subject. These spoofed images are then evaluated using cosine similarity against the true embeddings of the victims' iris images. A successful spoof occurs when the cosine similarity exceeds the optimal threshold of 0.6239, as previously derived from the similarity distributions in Figure 5.2.

These results demonstrate that our generator is highly effective at generating spoofed images

Table 5.1: Summary of attack success rates (ASR) and spoofed identity counts across 197 attackers against the Classification-Based Recognizer.

Metric	Mean	Min	Max
Sample-Level ASR	0.906	0.397	0.993
Subject-Level ASR	0.973	0.404	1
Victims Spoofed (count)	187.360	80	196

that fool the classification-based recognizer, even under strict similarity thresholds. As seen in Table 5.1, on average, attackers spoofed over 90% of individual samples, and more than 97% of victim identities were compromised through majority-vote attacks. Many attackers reached a 100% Subject-level ASR across all targets.

This high ASR reflects the idealized conditions of a white-box scenario, where the generator benefits from direct feedback during training via identity-preserving loss terms. While this serves as a good upper bound for attack performance and gives promising results, it doesn't necessarily mean that the model generalizes to unseen recognition systems. In the next section, we evaluate our attack under a more realistic black-box scenario using the HDBIF matcher, which represents a traditional recognition pipeline.

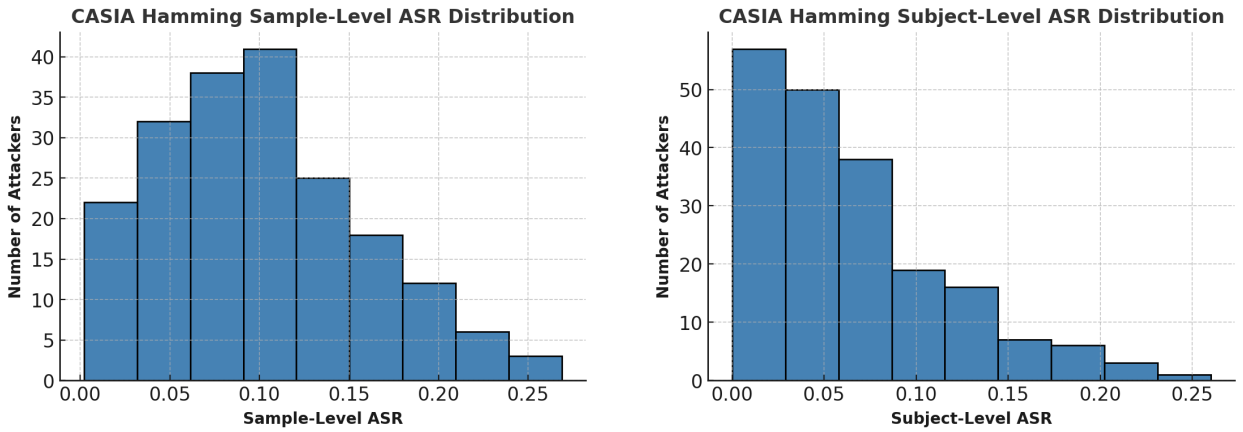
Black-Box ASR (Hamming Distance)

Similar to our white-box approach, to evaluate spoof success, we treat each subject in the filtered CASIA test set (197 subjects total) as an attacker in turn. All 197 subjects within this attack evaluation are unseen during the training of both the similarity model as well as the generator, ensuring that the results reflect subject generalization and true adversarial performance under black-box conditions.

Despite the ideal conditions of the CASIA dataset, the attack success rates under Hamming distance remain relatively low. As seen in 5.2, an average of only 6.4% of spoofed images

Table 5.2: Summary of attack success rates (ASR) and spoofed identity counts across 197 attackers against the HDBIF matcher.

Metric	Mean	Min	Max
Sample-Level ASR	0.064	0.000	0.269
Subject-Level ASR	0.042	0.000	0.260
Victims Spoofed (count)	12.416	0	51



(a) Sample-Level ASR distribution

(b) Subject-Level ASR distribution

Figure 5.5: Distribution of attack success rates (ASR) across all 197 attackers in the CASIA dataset. Sample-level ASR measures the proportion of successful spoofed images. Subject-level ASR reports the percentage of victim identities for which the attacker succeeds in a majority of trials.

bypass the hamming distance metric. Moreover, attackers were only able to able to spoof an average of 4.2% of victims through majority vote. While some attackers achieved over 25% of targets, the overall results suggest that traditional binary-code-based iris matches are resilient to this type of digital presentation attack.

5.2.3 Discussion

The CASIA-based evaluation reveals a significant contrast in ASR values between white and black-box scenarios. Spoofing performance is high in the white-box setting, where the generator receives identity-level feedback from the same recognizer used for evaluation. On

average, attackers successfully impersonate over 90% of spoofed samples and 97% of victim identities. This reflects the nature of the white-box setup and validates the generator's ability to preserve identity cues during the swap process.

On the other hand, when this model is evaluated under black-box conditions using hamming distance as a metric, spoofing effectiveness drops sharply. Average sample-level ASR falls to just 6.4%, and subject-level ASR averages only 4.2%. These results suggest that iris recognition pipelines using binary encoding and Hamming distance as a metric remain resilient against this attack pipeline when it's trained solely on a CASIA dataset.

However, a closer look reveals that a subset of attackers still achieve moderate spoofing success. Some subjects were able to spoof over 25% of the victim population, indicating that specific iris textures or segmentation artifacts can occasionally align favorably with binary matching logic. While the average performance remains low, these isolated successes highlight potential vulnerabilities that could be exacerbated with further tuning, better segmentation, or Hamming-specific optimization.

This performance gap also reflects a key limitation of our model: while the generator is optimized using cosine similarity in a continuous embedding space, the black-box matcher relies on a fundamentally different representation (binary iris codes). As a result, our identity-guided training does not directly translate into Hamming distance spoofing success. Meaning, our identity-guided training does not allow adversarially generated images to bypass the Hamming Distance metric indirectly.

Finally, the CASIA dataset itself imposes limitations. While it offers high-quality, near-infrared iris images under controlled settings, it lacks the variability seen in real-world or VR-captured imagery. This generalizes more realistic conditions, less certain, and motivates further experiments on OpenEDS and Neon datasets.

The CASIA dataset itself also imposes notable constraints. While it offers high-resolution, near-infrared iris images captured under controlled settings, it lacks the variability found in real-world scenarios or VR-based data. Moreover, our data filtering procedure, which discards poorly segmented samples and requires a minimum number of valid samples per identity, results in a substantially reduced dataset size. After filtering, we retain only about 5,000 usable images. This reduced scale limits the diversity of training samples and may contribute to overfitting or poor generalization to unseen recognition systems.

Together, these findings highlight the importance of evaluating attack robustness under multiple recognition frameworks and across datasets with greater variability. This motivates our subsequent experiments on the OpenEDS datasets, which provide more challenging and realistic testing environments.

5.3 OpenEDS2019

The OpenEDS2019 dataset introduces more of a realistic and challenging domain for iris-based authentication. Collected using a head-mounted eye-tracking device, this dataset includes eye images captured under natural ambient lighting with variation within the gaze directions. It also includes a range of partial eye closers and eyelid occlusions. Each image is labeled with subject identity and comes with a high-quality binary iris segmentation mask indicating the iris region, removing the reliance for an external segmentation model.

In this section, we will evaluate the generalizability of our attack pipeline in a VR domain-specific dataset and its generalizability to unseen individuals both within and outside of the OpenEDS2019 domain.

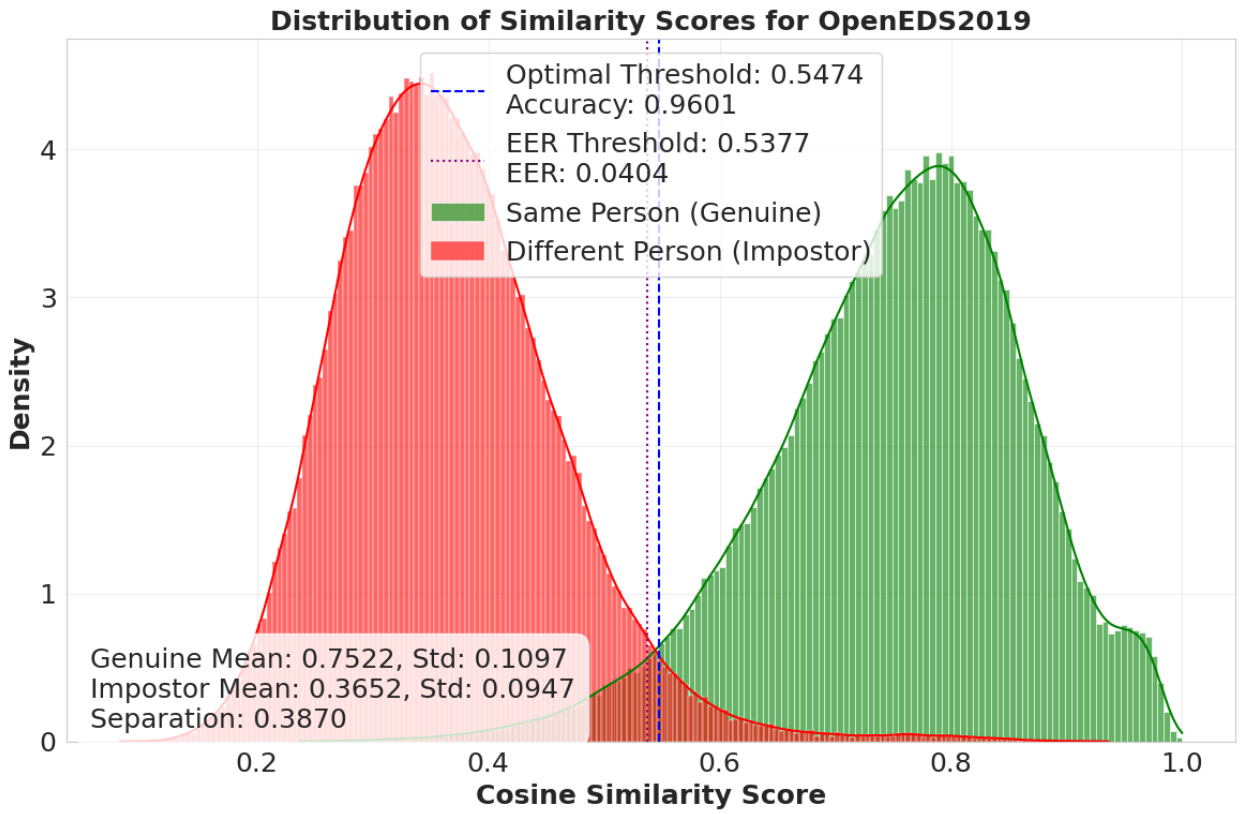


Figure 5.6: Distribution of cosine similarity scores between genuine and impostor pairs on unseen individuals in the OpenEDS2019 dataset using the classification-based recognizer. We sample up to 30 same-class and different-class comparisons per image to ensure a representative distribution of each image from unseen individuals. The two distributions show high separation, with an EER value of 4.04% with an accuracy of 96.01%.

5.3.1 Identity Recognition Performance

Classification-based Recognizer

We begin our evaluation by examining the effectiveness of the classification-based recognizer trained solely on the OpenEDS2019 dataset, identifying whether or not this similarity model methodology can generalize to other domains. Since this dataset contains monocular images, labels are set at the subject level during training. Images with no identified iris region, such as blinks, were removed during the training and evaluation of this similarity model. Figure 5.6

shows the cosine similarity distributions of genuine and imposter iris pairs within unseen subjects within the OpenEDS2019 dataset. The model achieves an EER value of 4.04% with an accuracy of 96.01%. We set the optimal threshold to be at 0.5474 based on classification accuracy.

Despite the large variation seen within OpenEDS2019, the classification-based recognizer effectively distinguishes between real identities and provides a reliable reference point for evaluating spoofed images generated by our swap model.

Recognizer Generalization to OpenEDS2020

To test cross-VR-domain generalizability, we evaluate the same OpenEDS2019-trained classification-based recognizer on OpenEDS2020. In addition to featuring entirely different subjects, OpenEDS2020 introduces variation in lighting conditions, creating a meaningful domain shift from the original training data.

As shown in Figure 5.7, the separation between genuine and imposter distributions is reduced. The Equal Error Rate increases to 28.47%, and the classification accuracy drops to 72.95%. This decline in performance highlights the limited cross-domain robustness of the OpenEDS2019-trained similarity model, particularly when evaluated on images with different eye shapes, illumination patterns, or capture conditions.

These findings motivate the experiments in the next section, where we train across both OpenEDS2019 and CASIA domains to improve generalization to unseen VR datasets such as OpenEDS2020.

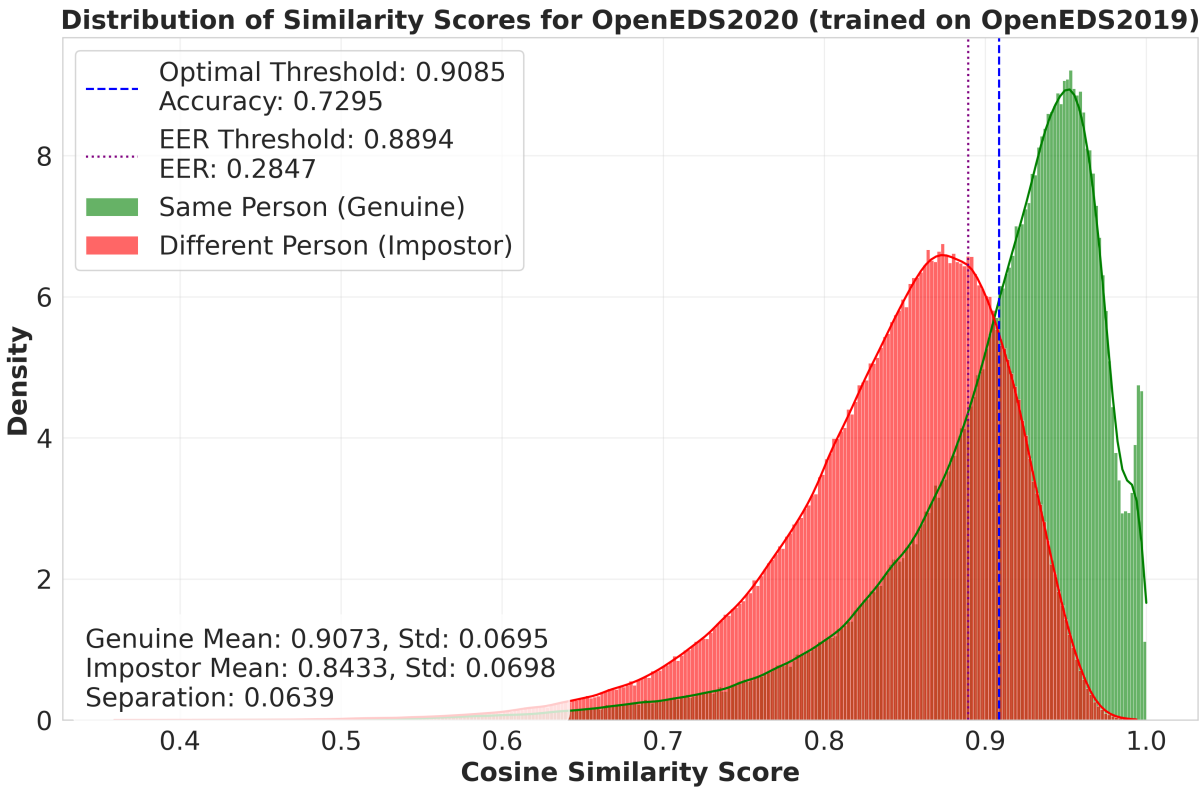


Figure 5.7: Cosine similarity distributions for genuine and impostor pairs when evaluating the OpenEDS2019-trained recognizer on OpenEDS2020. High overlap (EER value of 28.47%) between the distributions indicates weaker generalization performance to unseen datasets.

Hamming Distance Comparisons

Unlike the CASIA dataset, OpenEDS2019 contains off-axis gaze angles, partial eyelid closures, and increased variability in eye orientation due to its capture in a head-mounted display environment. These factors reduce the reliability of traditional Hamming distance–based matching. In contrast, CASIA consists of high-resolution, frontal, near-infrared images with minimal variation in gaze or lighting. These properties align well with the assumptions of the Hamming distance framework and enable strong separation between same-subject and different-subject comparisons, as shown in Section 5.2.1. Although OpenEDS2019 is somewhat ill-suited for traditional iris recognition pipelines, such as Hamming distance matching, we perform an exploratory evaluation using this metric for completeness.

Hamming distance requires consistent iris alignment and high-contrast binary iris codes to measure similarity accurately. However, images in OpenEDS2019 often include occlusions, conditions that violate the assumptions of classical iris code generation. As a result, we observed significant overlap between genuine and impostor score distributions when evaluating OpenEDS2019 with the HDBIF matcher, leading to a high equal error rate (EER).

As seen in Figure 5.8, the equal error rate is at 26.88% for the unseen 31 individuals within the test set, with an accuracy rate of around 73%. This indicates a much weaker separation of the two genuine and imposter distributions than the CASIA figure. We set the threshold to be at 0.4134 to maximize the classification accuracy between pairwise comparisons.

Contrastive Recognizer Limitations

Similar to our CASIA results in Section 5.2.1, we also experimented with a contrastive recognizer using InfoNCE. However, the limited dataset size and subject diversity in OpenEDS2019 hindered its effectiveness. While the dataset contains many eye images, the number of

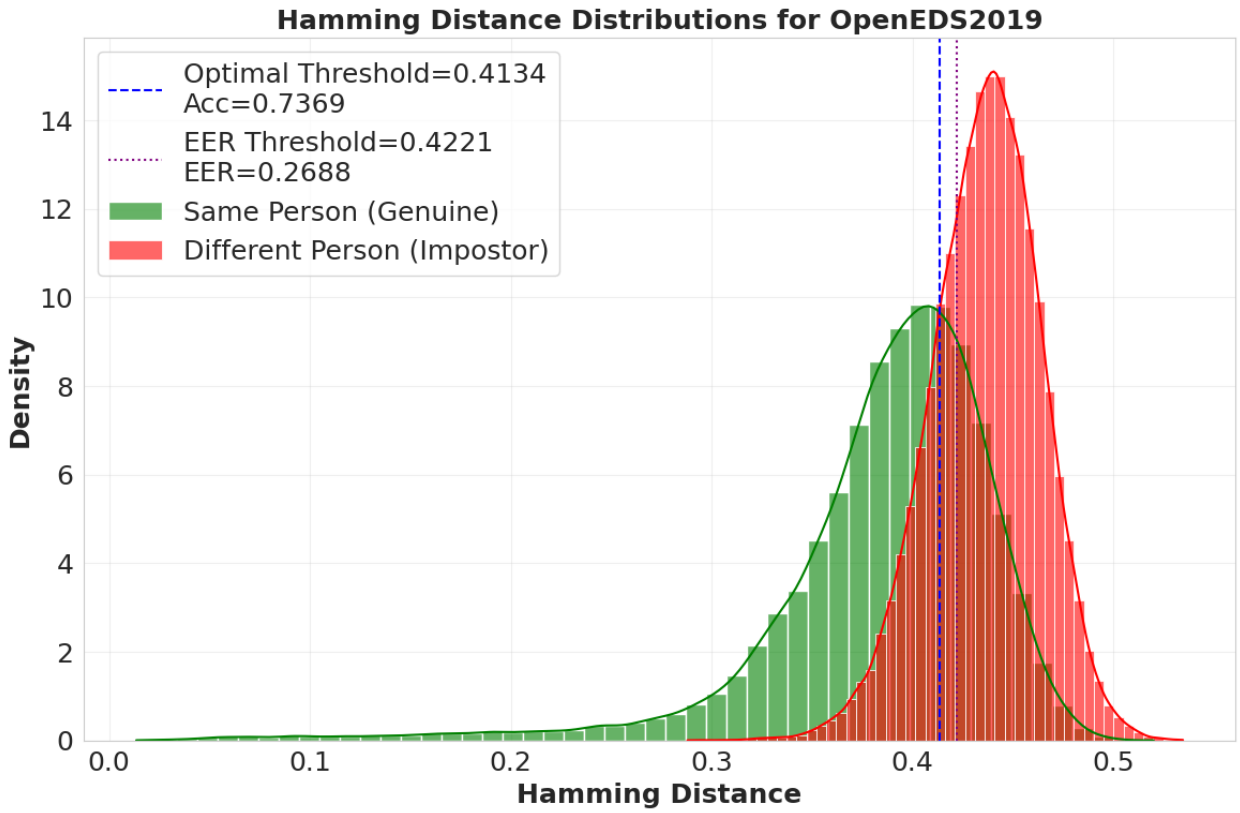


Figure 5.8: Distribution of Hamming distances between genuine and impostor pairs in the OpenEDS2019 dataset using the HDBIF matcher. The distributions exhibit high overlap with an EER value of 26.88%, and an accuracy of 73.69%.

unique identities (152 subjects) and limited variation per subject constrained the diversity of anchor-positive-negative pairs required by contrastive learning. As a result, the learned embedding space failed to generalize, exhibiting substantial overlap between genuine and impostor distributions.

This outcome is consistent with findings in contrastive learning literature, which emphasize the need for large-scale, diverse datasets to prevent embedding collapse and sufficiently learn identity-separating features. Due to this limitation, we did not continue using the contrastive recognizer in further OpenEDS2019-specific evaluation. We examine the effectiveness of contrastive learning of a multi-domain training set within Section 5.4

5.3.2 Swap-Based Attack Evaluation

To evaluate the effectiveness of our methodology, we trained our GAN architecture model on the OpenEDS2019 dataset. Similar to our CASIA evaluation, we evaluate attack effectiveness under two different settings:

- **White-box Attack:** Victim and generated image pairs are input into the contrastive recognizer and evaluated against the optimal threshold identified by the maximum pair-wise accuracy. Since the generator received feedback from this model during training, this represents a best-case scenario for spoofing success.

- **Black-box attack:** Victim and generated image pairs are input into the HDBIF matcher to generate Hamming Distance values for each image pair. These values are then evaluated against the identified optimal threshold.

Training Setup

Following a similar procedure to CASIA, we apply a data filtering process to the OpenEDS2019 dataset to ensure high-quality input into the generator. Because this dataset contains monocular images with greater variability in eye openness, we exclude samples with no identified iris region, such as frames with blinks or occlusions. Unlike CASIA, eye-specific labels are not available. Therefore, the labels are at the subject level during training.

Evaluation Setup

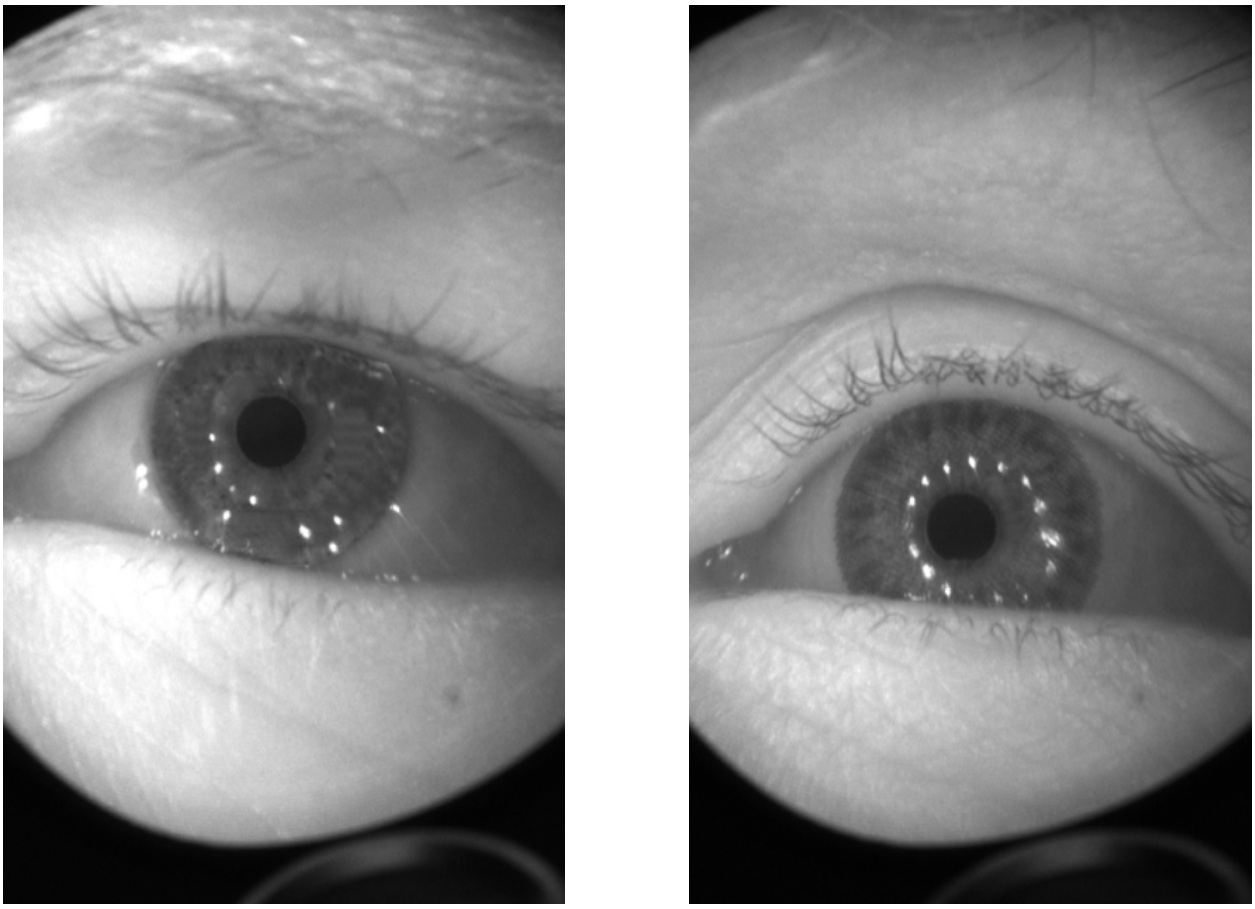
Our spoof evaluation protocol for OpenEDS2019 mirrors the CASIA setup described in Section 5.2.2, with adjustments for VR-specific image variability. Each subject in the test set acts as an attacker, attempting to spoof every other subject using generated images. The victim's iris is chosen as the sample with the largest segmented iris region, using the same area-maximization strategy defined for CASIA:

As with CASIA, all spoofed image pairs are constructed using test-time subjects that are completely disjoint from those seen during training, ensuring robust evaluation. An example of a spoofed and victim pair from OpenEDS2019 is shown in Figure 5.9.

White-Box ASR (Classification-Based Recognizer)

To evaluate the effectiveness of our attack within a white-box environment, where the adversary has full knowledge of the target system's internals. In our case, the generator was receiving feedback from our classification-based recognizer. Therefore, it is inherently optimized to bypass it. This represents the most favorable conditions for attack success rate.

Following the same evaluation protocol described earlier, we are left with 31 subjects within



(a) Spoofed image (b) Victim image

Figure 5.9: Example spoofed and victim image pair from the OpenEDS2019 dataset. The spoofed image was synthesized by inpainting the victim's iris into the attacker's eye region.

our evaluation set (out of 152). Each subject within the evaluation set is an attacker for all the other 30 subjects to compute our results. Similarity values between attacker and victim frames above 0.5474 are considered a successful spoof.

As shown in Table 5.3, all attackers could successfully spoof every other target identity within the evaluation set, achieving a perfect subject-level ASR of 1.0. Similarly, the sample-level ASR is high with a mean of 0.997. These results highlight the strength of our generator under white-box conditions, when it is directly optimized against the same model used for evaluation.

Table 5.3: Summary of near-perfect attack success rates (ASR) of unseen OpenEDS2019 subjects against the Classification-Based Recognizer trained on OpenEDS2019 data. This high ASR indicates that our methodology works in white-box settings against our recognizer.

Metric	Mean	Min	Max
Sample-Level ASR	0.997	0.950	1
Subject-Level ASR	1	1	1
Victims Spoofed (count)	30	30	30

Black-Box ASR (Hamming Distance)

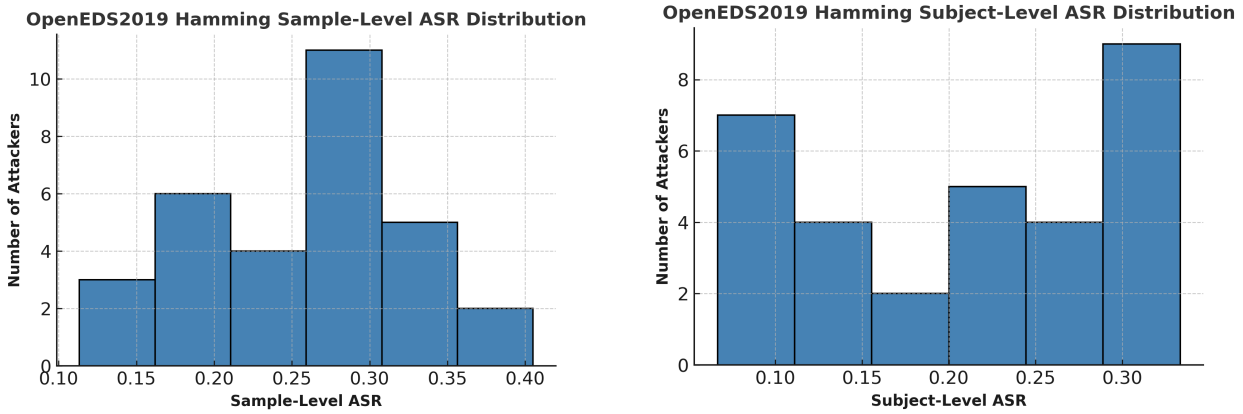
While our generator is trained to optimize cosine similarity within a continuous embedding space, Hamming distance offers a fundamentally different matching paradigm. It is not used during training and, therefore, treated as a black-box recognizer. This setup simulates an adversary trying to bypass a real-world matcher without access to its internal representation, parameters, or gradients.

In this evaluation, all 31 subjects are unseen during the training of the Classification-Based Recognizer and the generator. We classify any victim & generated image pair below the optimal threshold set at 0.4134 to be a successful spoof.

Table 5.4: Summary of attack success rates (ASR) of the 31 unseen OpenEDS2019 subjects against the Hamming distance threshold of 0.4134. The meaningful, yet low, ASR scores indicate that the influence of our recognizer model doesn't indirectly bypass the black-box hamming distance metric.

Metric	Mean	Min	Max
Sample-Level ASR	0.262	0.113	0.405
Subject-Level ASR	0.210	0.067	0.333
Victims Spoofed (count)	6.29	2	10

Table 5.4 shows that our generator achieves high ASR against the Hamming distance threshold. Among 31 evaluated subjects, over 25% of spoofed images were accepted on average, with multiple attackers successfully impersonating 30-33% of possible victims through majority voting.



(a) Sample-Level ASR distribution

(b) Subject-Level ASR distribution

Figure 5.10: Distribution of attack success rates (ASR) across all 31 attackers in the OpenEDS2019 dataset. Sample-level ASR measures the proportion of successful spoofed images. Subject-level ASR reports the percentage of victim identities for which the attacker succeeds in a majority of trials.

5.3.3 Discussion

The OpenEDS2019 results reveal that our attack pipeline performs exceptionally well under white-box conditions, achieving perfect subject-level spoofing success. This confirms that when the generator is optimized directly against the classification-based recognizer, spoofed images can mimic real identities, even within the challenging conditions of VR image domains. These conditions include partial eyelid occlusions, gaze variability, and lighting changes, which were absent in CASIA.

However, the black-box results reveal a significant drop in performance. Although still stronger than CASIA Hamming distance results, the average spoof success rate using Hamming distance was 26.2% at the sample level and 20.9% at the subject level. This disparity underscores a key insight: while our generator can optimize for cosine similarity, it struggles to generalize to binary-code-based matches that were not part of its training process. Hamming distance operates under a fundamentally different matching paradigm, and OpenEDS2019's noisier data introduces more significant variability in iris code generation,

making spoofing success inconsistent.

Hamming distance relies on a fundamentally different matching paradigm, and the noisier, off-axis imagery in OpenEDS2019 introduces greater variability into iris code generation, making spoofing success inconsistent. Notably, the Hamming threshold used for OpenEDS2019 in this evaluation (0.41) is considerably more permissive than thresholds typically recommended in the literature (around 0.38), which may have inflated attack success. As shown later in our mixed-domain experiments, when the Hamming threshold is tightened to 0.38, spoofing success rates decrease further and better align with prior biometric security findings [8, 14].

Nevertheless, this level of success should not be dismissed. Among 30 evaluated subjects, several attackers still managed to spoof 30–33% of victims via majority vote (Table 5.4, Figure 5.10). These results suggest that when the matcher itself is unreliable, even weakly aligned spoofed images can bypass recognition, particularly in the domain of VR images. Rather than reflecting the generalization of the swap model, this highlights the fragility of outdated matchers in unconstrained VR environments.

Furthermore, our contrastive recognizer failed to generalize effectively on OpenEDS2019. Although the dataset includes a high volume of images, the relatively low number of subjects and limited intra-subject diversity reduce the effectiveness of contrastive learning. This reinforces the understanding that contrastive methods require larger and more diverse identity sets to succeed in biometric recognition tasks.

These findings motivate our next set of experiments, which explore whether training across both CASIA and OpenEDS2019 can improve spoofing performance in mixed-domain scenarios and enhance generalization to entirely new datasets such as OpenEDS2020.

5.4 CASIA + OpenEDS2019

To evaluate whether increasing dataset diversity can improve generalization in iris-swapping attacks, we train our full pipeline on a combined datset composed of CASIA and OpenEDS2019. This setup allows us to explore how models trained on both traditional near-infrared iris images and VR-captured imagery perform in both seen and unseen domains. Compared to previous experiments, this setup introduces a more challenging training distribution due to domain shifts in image quality.

To tackle this, we adopt a contrastive learning strategy for our iris recognizer to learn a unified embedding space that can span both domains. This section evaluates whether a multi-domain recognizer trained with InfoNCE loss improves generalization over prior classification-based approaches.

We begin by evaluating attack success rates on in-domain samples from both CASIA and OpenEDS2019. Then, we assess generalization to an entirely unseen dataset: OpenEDS2020. In this evaluation, OpenEDS2020 is unseen to both the recognizer and the GAN swap model.

Lastly, we do a Hamming distance evaluation on CASIA data for thoroughness for our black-box scenario. Here, we quantify whether the influence of our contrastive recognizer improves on a Hamming distance metric unseen to the models during training.

5.4.1 Contrastive Recognizer

To overcome the limitations of classification-based recognition, particularly in open-set and cross-domain evaluation settings, we train a domain-generalizable iris recognizer using contrastive learning. Specifically, we use the InfoNCE loss to learn an embedding space where images from the same individual cluster close together, while those from different individuals

are pushed apart.

For training, we combine samples from both CASIA and OpenEDS2019, ensuring a diverse representation of traditional near-infrared iris images and VR-captured images. Within this combined training set, we reserve a disjoint subset of individuals from each domain for evaluation to ensure fair measurement of generalization.

We evaluate this recognizer under three conditions:

- **Within-Domain: CASIA** — Evaluation on CASIA test subjects unseen during training.
- **Within-Domain: OpenEDS2019** — Evaluation on OpenEDS2019 test subjects.
- **Cross-Domain: OpenEDS2020** — Evaluation on the entirely unseen OpenEDS2020 dataset to assess cross-dataset generalization.

Within domain: CASIA

Following a similar evaluation protocol as previous Identity Recognition Sections, we compute similarity distributions between images from unseen individuals within the evaluation set. Shown in Figure 5.11, for the 197 CASIA subjects that were unseen by the training of this recognizer, the model achieves an EER value of 3.38% with an accuracy value of 97.88%. We set the optimal threshold for white-box within domain attacks to be 0.8101 based of pairwise classification accuracy.

This result highlights that, even when trained jointly on mixed-domain data (CASIA + OpenEDS2019), the contrastive framework is capable of preserving high identity separability on traditional near-infrared iris images. The tight separation between genuine and impostor distributions suggests that contrastive learning successfully embeds identity-specific features

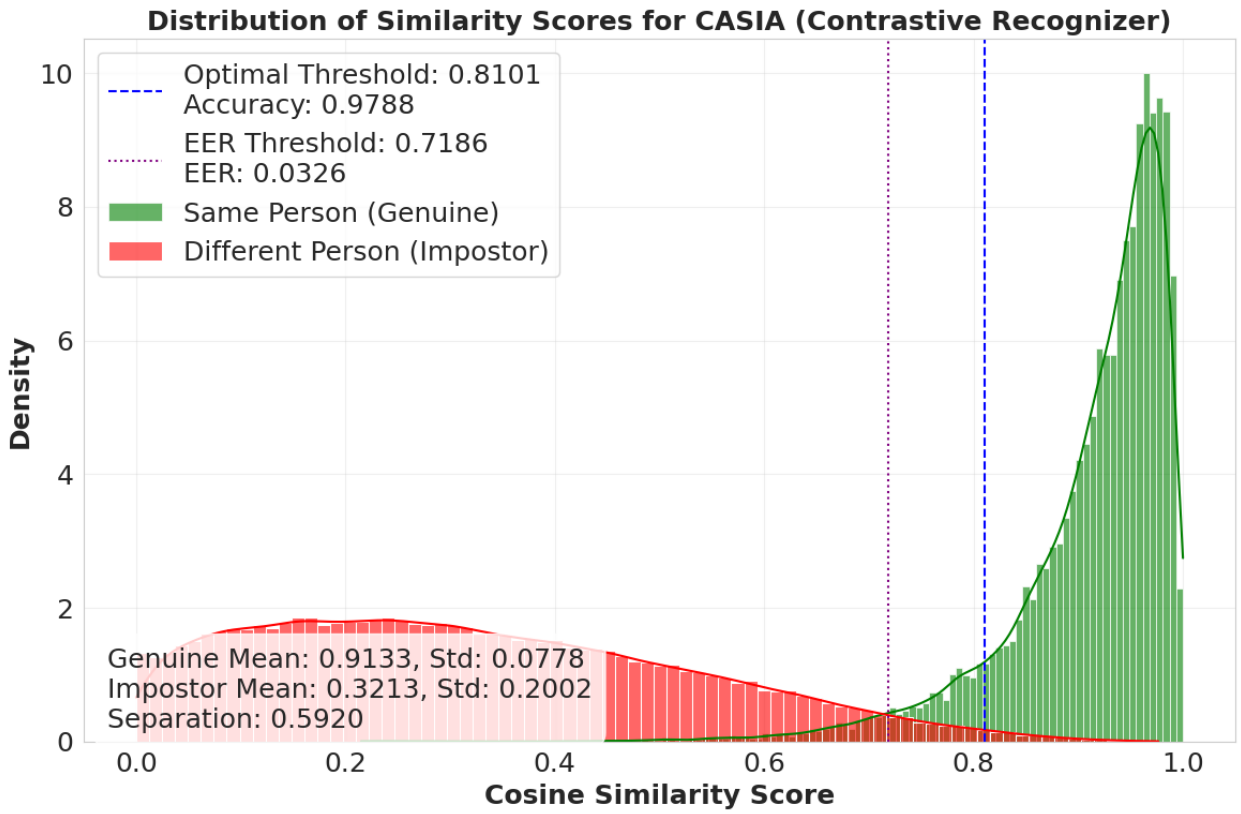


Figure 5.11: Distribution of Cosine Similarity values from image pairs from the 197 unseen CASIA subjects. 30 same-class and different-class comparisons per image were taken to ensure a representative distribution. This graph indicates a low distribution overlap between same-class & different-class distributions, with a low EER rate of 3.26%. This indicates that our recognizer model performs well on unseen CASIA individuals.

while generalizing across domain shifts.

Within domain: OpenEDS2019

We now evaluate the contrastive recognizer on 31 held-out OpenEDS2019 subjects that were excluded from training. As shown in Figure 5.12, the model achieves an Equal Error Rate (EER) of 13.74% and an accuracy of 86.26%. While performance is lower compared to CASIA, the contrastive model maintains a clear margin between genuine and impostor pairs despite the greater visual variability in VR-captured data. We set the optimal cosine

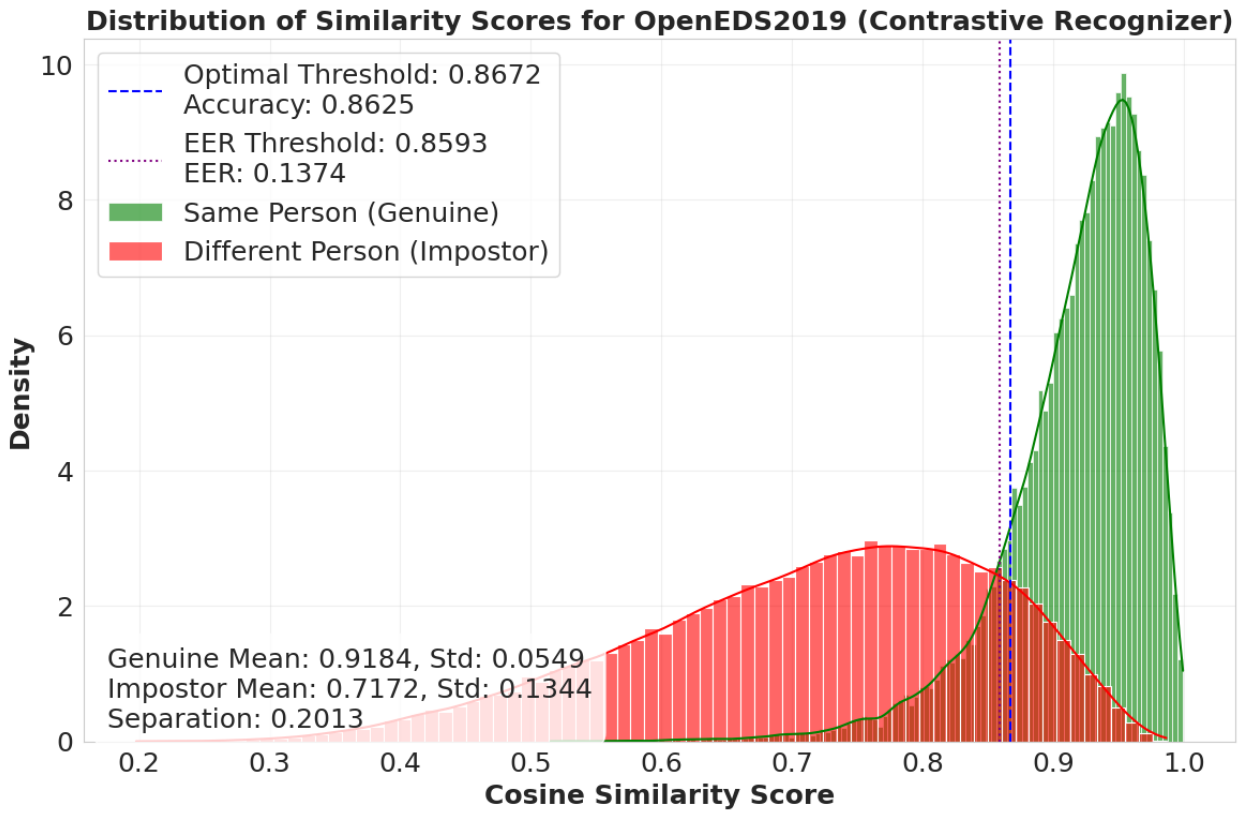


Figure 5.12: Distribution of Cosine Similarity values of image pairs from the 31 unseen OpenEDS2019 subjects. 30 same-class and different-class comparisons per image were generated to ensure a representative distribution. The relatively low EER (13.74%) indicates that the recognizer can differentiate between genuine and imposter pairs of irises.

similarity threshold for white-box attacks in this domain to 0.8672.

Cross-Domain: OpenEDS2020

To test the generalization of our contrastive recognizer beyond the training domains, we evaluate its performance on OpenEDS2020. This VR dataset introduces new subjects and lighting conditions unseen in the other datasets. The results in Figure 5.13 demonstrate that the model maintains strong identity separation despite the domain shift. We achieve an EER of 8.47% and an overall accuracy of 91.61%, with a clear separation of 0.3712 between

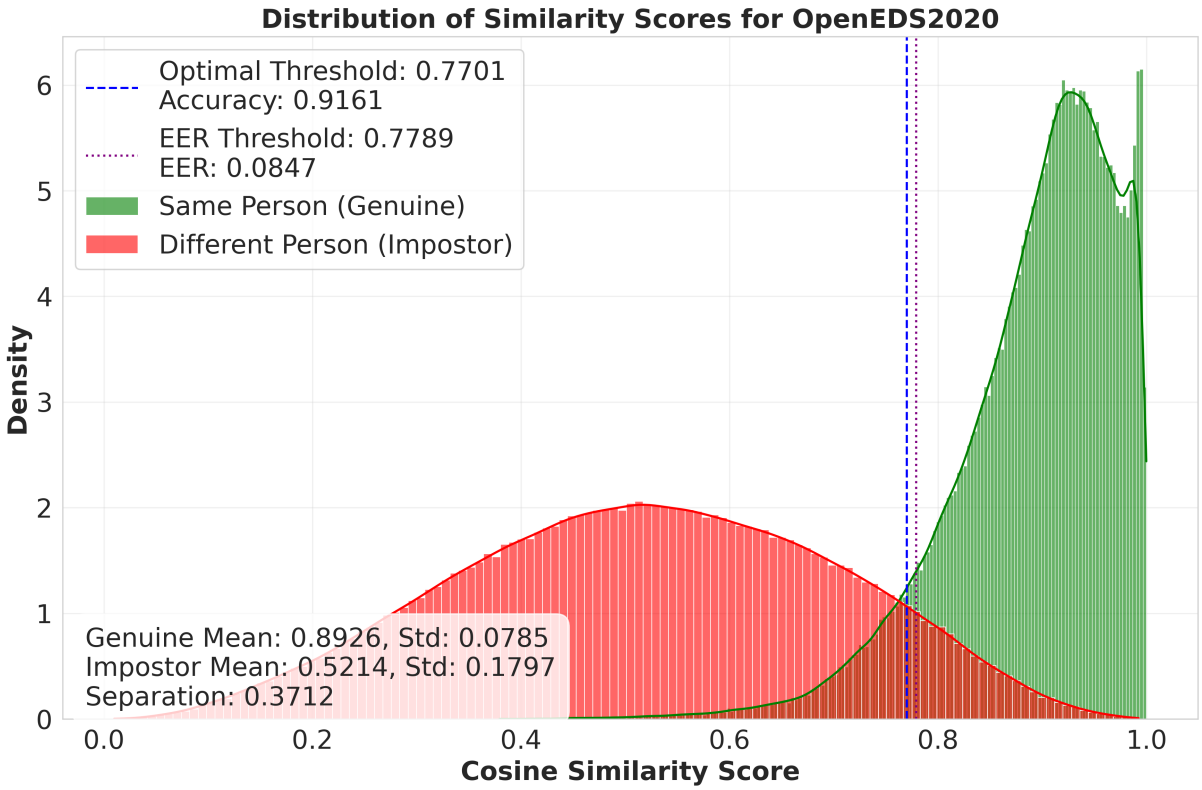


Figure 5.13: Distribution of cosine similarity values for image pairs from OpenEDS2020, using the contrastive recognizer trained jointly on CASIA and OpenEDS2019. Each unseen subject's image was used in 30 genuine and impostor pairs to evaluate cross-domain generalization. The low overlap between the two distributions indicates high generalization of our recognizer to unseen datasets.

the mean cosine similarity of genuine and impostor pairs.

These findings suggest that contrastive training across multiple domains fosters better generalization, allowing the model to scale to new VR datasets without additional fine-tuning. In particular, the low EER is noteworthy given that OpenEDS2020 includes subjects entirely disjoint from the training set within a different domain. This reinforces the benefit of domain-diverse training for improving robustness to unseen data distributions in biometric applications.

5.4.2 Swap-Based Attack Evaluation

This section evaluates the attack success of our iris-swapping pipeline across different recognition settings and dataset domains. While previous sections established baseline identity recognition performance, here we investigate whether our generator can successfully produce spoofed images that impersonate unseen identities.

All spoofed image evaluations in this section are conducted using a contrastive learning-based recognizer trained jointly on the CASIA and OpenEDS2019 datasets. This model allows for generalizable identity matching across domains and is the primary metric for evaluating the success of white-box attacks.

We test spoofing effectiveness under both white-box and black-box assumptions:

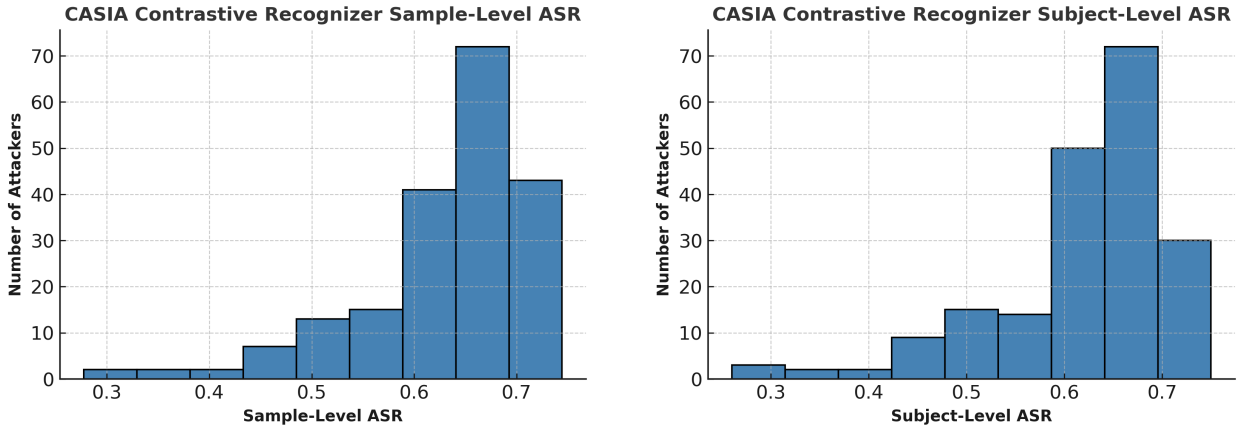
- **White-box (CASIA)**: The generator is evaluated against the same classification-based recognizer it was optimized with during training. This setting offers an upper bound on spoofing effectiveness, since the generator receives direct gradient feedback from the recognizer.
- **White-box (OpenEDS2019)**: We perform a similar evaluation using a model trained entirely within the VR domain, showing spoofing success under realistic imaging conditions such as varied gaze angles and partial occlusions.
- **Attack Generalization (OpenEDS2020)**: To evaluate attack generalization, we generate OpenEDS2020 spoofs using a GAN model influenced and training on CASIA+OpenEDS2019. The goal is to evaluate the transferability of spoofs to unseen subjects and datasets.
- **Black-box (CASIA Hamming Distance)**: Finally, we evaluate spoofed CASIA images using the HDBIF matcher, a traditional binary-code-based iris recognizer. This

setting assesses the model's ability to generalize beyond the continuous feature space used during training.

White-box ASR: CASIA

Table 5.5: Summary of attack success rates (ASR) of the 197 unseen subjects against the Contrastive Recognizer trained on both the CASIA & OpenEDS2019 datasets. In both sample-level and subject-level ASR, we're able to spoof the recognizer in more than half of our samples.

Metric	Mean	Min	Max
Sample-Level ASR	0.632	0.278	0.744
Subject-Level ASR	0.622	0.260	0.750
Victims Spoofed (count)	121.99	51	147



(a) Sample-Level ASR distribution

(b) Subject-Level ASR distribution

Figure 5.14: Attack Success Rate (ASR) distributions for the 197 unseen CASIA subjects against the contrastive recognizer trained on CASIA and OpenEDS2019. Sample-level ASR reflects the proportion of spoofed images deemed successful. Subject-level ASR measures the percentage of victims for which spoofing succeeded via majority vote.

As shown in Table 5.5, the generator achieves an average sample-level ASR of 63.2% and a subject-level ASR of 62.2%. On average, attackers successfully spoofed over 121 out of 196 potential victims. The distribution plots in Figure 5.14 further illustrate this trend.

Most attackers achieved subject-level success rates between 60% and 75%, indicating broad transferability across identities.

White-box ASR: OpenEDS2019

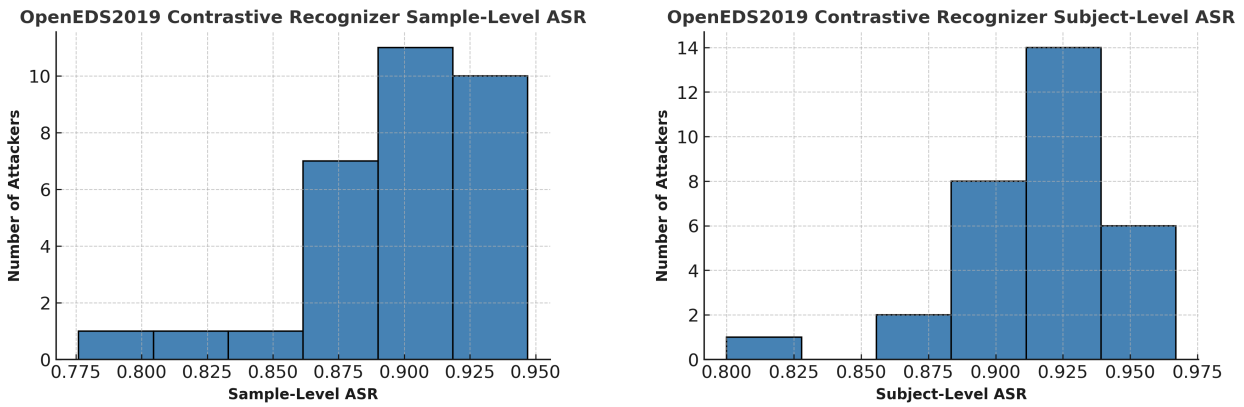
To assess white-box spoofing performance within a VR-specific domain, we evaluate our generator against the contrastive recognizer trained jointly on CASIA and OpenEDS2019. All 31 evaluation subjects were withheld from training and treated as attackers attempting to spoof each of the other 30 subjects. The decision threshold of 0.8593, derived from OpenEDS2019 similarity distributions, determines spoofing success.

Table 5.6: Summary of attack success rates (ASR) of the 31 unseen subjects against the Contrastive Recognizer trained on both the CASIA & OpenEDS2019 datasets. These high ASR's indicate that our swap model performs well in white-box settings for unseen OpenEDS2019 individuals.

Metric	Mean	Min	Max
Sample-Level ASR	0.906	0.776	0.947
Subject-Level ASR	0.925	0.800	0.967
Victims Spoofed (count)	27.67	24	29

Table 5.6 summarizes the ASR performance: attackers achieved an average sample-level ASR of 90.6% and subject-level ASR of 92.5%, indicating strong generalization and spoofing capability within the OpenEDS2019 domain.

Figure 5.15 illustrates the distribution of ASR scores across all attackers. Most attackers spoofed nearly every victim in the test set, with subject-level ASR clustering above 90%. These results demonstrate that the contrastive model preserves spoofing susceptibility in VR data even under domain-aware identity learning.



(a) Sample-Level ASR distribution

(b) Subject-Level ASR distribution

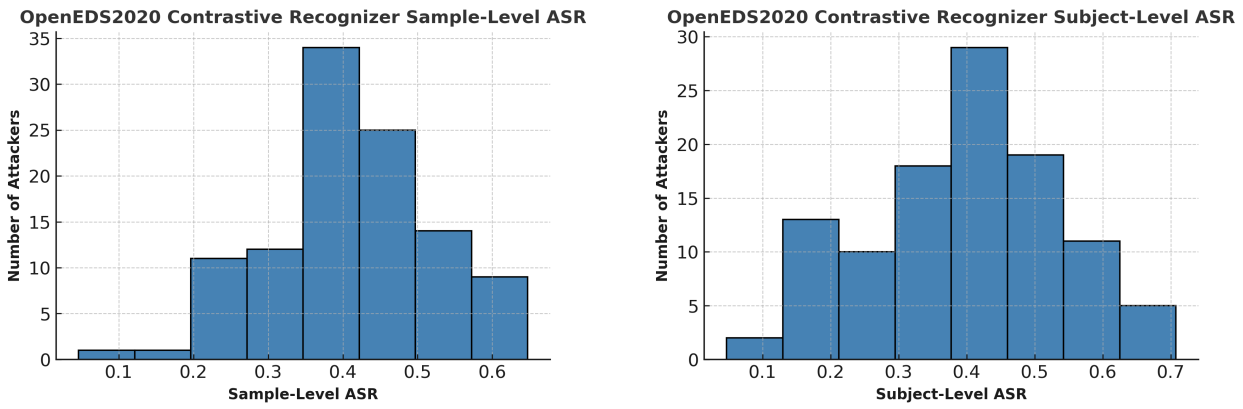
Figure 5.15: Attack Success Rate (ASR) distributions for the 31 unseen OpenEDS subjects against the contrastive recognizer trained on CASIA and OpenEDS2019. Sample-level ASR reflects the proportion of spoofed images deemed successful. Subject-level ASR measures the percentage of victims for which spoofing succeeded via majority vote.

Table 5.7: Summary of the black-box attack success rates (ASR) for the 107 subjects within OpenEDS2020 against the Contrastive Recognizer trained on CASIA & OpenEDS2019 datasets.

Metric	**Mean**	**Min**	**Max**
Sample-Level ASR	0.411	0.046	0.647
Subject-Level ASR	0.397	0.047	0.708
Victims Spoofed (count)	20.7	0	57

Attack Generalization: OpenEDS2020

The OpenEDS2020 results evaluate the cross-dataset spoofing effectiveness of the generator when assessed against a contrastive recognizer trained solely on CASIA and OpenEDS2019. As shown in Table 5.7, the mean subject-level ASR is 39.7%, with a maximum spoofing success of over 70% for some attackers. These results are particularly significant given that the generator and recognizer were trained without access to any OpenEDS2020 identities. The sample-level histogram in Figure 5.16a shows that the majority of attackers achieved spoofing success above the 40% mark, while the subject-level distributions in Figure 5.16b reveal a moderate spread in spoofability across the population. While not as high as same-



(a) Sample-Level ASR distribution

(b) Subject-Level ASR distribution

Figure 5.16: ASR distributions for 107 OpenEDS2020 subjects against the contrastive recognizer. Sample-level ASR reflects individual image spoofing success, while subject-level ASR indicates percentage of spoofed victim identities.

dataset results, the ASR in this dataset domain demonstrates meaningful generalization to unseen domains.

Black-Box ASR: CASIA Hamming Distance

Table 5.8: Updated ASR results using Hamming distance comparisons against spoofed CASIA images generated from the model using the contrastive learning-based recognizer. Our contrastive learning-based recognizer indicates slight improvement from our previous model.

Metric	Mean	Min	Max
Sample-Level ASR	0.143	0.000	0.320
Subject-Level ASR	0.110	0.000	0.291
Victims Spoofed (count)	20.7	0	57

To evaluate whether our multi-domain training and contrastive learning strategy yielded any black-box robustness improvements, we re-ran the Hamming distance attack evaluation using the spoofed images generated by the latest model. As before, we treat the HDBIF matcher [1] as a black-box system and measure whether spoofed images fall below the decision threshold of 0.38 derived from genuine/impostor CASIA distributions.

Table 5.8 reports updated ASR values across all 197 unseen subjects. On average, sample-level ASR increased from 6.4% to 14.3%, and subject-level ASR improved from 4.2% to 11.0% compared to the baseline model trained on CASIA alone.

5.4.3 Discussion

The CASIA + OpenEDS2019 experiments demonstrate the benefits of incorporating multi-domain training and contrastive supervision for recognition and spoofing contexts. On the recognition side, our contrastive recognizer achieved strong within-domain performance on CASIA (EER = 3.26%). It maintained a meaningful separation margin on OpenEDS2019 (EER = 13.74%) despite its vast differences from the CASIA dataset. Most notably, it generalized well to an entirely unseen domain, OpenEDS2020, achieving an EER of 8.47%. This is a significant improvement from Figure 5.7, where the classification-based similarity model had much higher overlap between the two distributions.

On the attack side, spoofing success remained strong within both training domains. In CASIA (white-box), attackers successfully impersonated over 60% of victims on average—slightly lower than in classifier-based settings but still substantial. In OpenEDS2019 (white-box), spoofing was even more effective, achieving a subject-level ASR of 92.5%. The most compelling evidence of generalizability came from our OpenEDS2020: although the generator had never seen this dataset, spoofed images achieved a subject-level ASR of 39.7%, outperforming earlier CASIA spoofing attempts against the Hamming Distance matcher.

While these results fall short of the near-perfect spoofing success seen under classification-based white-box attacks, the reduction in ASR is both expected and meaningful. The contrastive recognizer learns generalizable embeddings by optimizing pairwise similarity, making it better suited for evaluating spoofing success across heterogeneous datasets. In contrast,

the classification-based recognizer is optimized for identity prediction using a classification head, which biases the learned features toward domain-specific class boundaries. Although this head is removed at test time, the underlying feature space remains specialized for the training distribution. As a result, it performs well within-domain but fails to generalize. The contrastive model, on the other hand, provides a more realistic and challenging recognition target—and the strong spoofing success it still permits highlights the robustness of our generator, even in the absence of task-specific alignment.

These findings collectively address **RQ2** and **RQ3** from the thesis: *Can a GAN-based iris-swapping method impersonate identities using known iris data?* and *Do iris-swapping attacks generalize across AR/VR datasets and biometric pipelines?* Our results show that they do—particularly when supported by contrastive learning—but with caveats. Generalization to unseen datasets is feasible, though spoofing effectiveness diminishes as the domain gap grows. This reinforces the need for diverse training distributions in both spoof detection and identity recognition models.

Chapter 6

Conclusions

In this thesis, we investigated vulnerabilities in iris-based biometric authentication within AR/VR systems, focusing on two attack vectors: biometric leakage from generative models and digital presentation attacks using GAN-based iris-swapping. We demonstrated that diffusion models fine-tuned on iris datasets could memorize and leak iris patterns of individuals within the training set. We also developed and evaluated a cross-attention GAN that achieved near-perfect spoofing success within dataset domains and meaningful generalization success across unseen VR datasets. Our findings highlight both the effectiveness and limitations of current generative attacks on biometric systems.

6.1 Summary of Findings

In Section 4.1, we demonstrated that Stable diffusion models trained on iris data can leak identities within their training sets. In Table 4.1, we show that when we fine-tune a Stable Diffusion model on 100 individuals within the CASIA-Iris-Thousand dataset, we were able to regenerate images that passed a hamming distance threshold of 0.38 when compared to the original training images. These findings provide direct evidence that generative models can memorize biometric information, even when fine-tuned on relatively small datasets, and highlight significant privacy risks for iris biometrics.

In Chapter 5, we evaluated the feasibility of digital iris-swapping attacks using a Cross-

Attention GAN architecture. We first trained and validated our models on the CASIA-Iris-Thousand dataset, achieving near-perfect spoofing success under white-box conditions. However, spoofing success against black-box Hamming distance matches remained limited, indicating that the cosine similarity optimization does not automatically translate into bypassing traditional biometric matches.

Next, we expanded our evaluation of our attack pipeline on the OpenEDS2019 dataset, simulating VR-specific capture conditions with off-axis gaze angles, eyelid occlusions, and light variations. Our experiments showed that although the generator maintained a near-perfect success rate under white-box conditions, our similarity model didn't generalize to other VR datasets. ASR also dropped when evaluating our model on the black-box Hamming distance conditions, further highlighting the challenge of generalizing spoof attacks across datasets and different matching pipelines.

Finally, we introduced a contrastive learning-based recognizer trained jointly on CASIA and OpenEDS2019, aimed to build a more domain-robust embedding space. Our experiments showed that:

- Identity separability remained strong within CASIA and OpenEDS2019 domains, maintaining low EERs despite increased variability.
- Swap-based attacks remained highly effective within domain, achieving subject-level ASR rates above 60% on CASIA and over 90% on OpenEDS2019.
- Reached meaningful cross-domain spoofing success: spoofed images were able to bypass the contrastive recognizer at a subject-level ASR of 39.7% on OpenEDS2020.

While white-box attack success against the contrastive recognizer was slightly lower than the earlier classification-based recognizer, this drop is both expected and meaningful. The

contrastive recognizer is trained to learn generalizable, domain-agnostic embeddings by optimizing pairwise similarity. On the other hand, the classification-based recognizer trained with a classification head learns decision boundaries tightly coupled to its training data. This limits generalization across domains and inflates white-box attack success rates.

These findings collectively answer the core research questions posed at the beginning of this thesis:

- **RQ1: Do generative models leak biometric training data?**
- **RQ2: Can a generative attack realistically spoof iris recognition systems?**
- **RQ3: Do iris-swapping attacks generalize across AR/VR datasets and biometric pipelines?**

6.2 Limitations

While this work demonstrates the feasibility of generative iris-swapping across VR datasets, several limitations remain. Although contrastive learning improved generalization across domains, spoofing success against our traditional Hamming distance metric remained low. Even with domain-robust embeddings, CASIA swap ASR remained relatively low (around 10-14%) against binary-code iris matchers. This suggests that the current training strategies do not fully bridge the gap between the embedding space of iris signatures and their binary biometric codes.

Second, this study focused on the offline evaluation of spoofed images. We did not implement or evaluate the real-time performance of this attack methodology into live camera streams within VR headsets. Exploring the real-time performance and practicality of such an attack

would be crucial to understanding the threat landscape for VR-based iris authentication systems.

6.3 Future Work

This work opens two promising directions for advancing research in biometric security.

First, future work could explore loss functions that explicitly optimize for binary-code similarity, such as differentiable Hamming approximations or hybrid training schemes that generate images with higher attack success rates against traditional matchers.

Second, expanding this attack pipeline to operate in a real-time condition remains an important next step. This would involve integrating the generator into a live camera feed on a VR headset operating in developer mode, enabling evaluation under real-world conditions and constraints.

Bibliography

[1] University of Notre Dame Open Source Iris Recognition Repository. https://github.com/CVRL/OpenSourceIrisRecognition/.

[2] Selfie biometrics: Advances and challenges. Springer International Publishing, 2019. ISBN 9783030269722. doi: 10.1007/978-3-030-26972-2. URL http://dx.doi.org/10.1007/978-3-030-26972-2.

[3] Fernando Alonso-Fernandez, Reuben A Farrugia, Josef Bigun, Julian Fierrez, and Ester Gonzalez-Sosa. A survey of super-resolution in iris biometrics with evaluation of dictionary-learning. *IEEE Access*, 7:6519–6544, 2018.

[4] Apple. Apple vision pro: Privacy and security. https://www.apple.com/apple-vision-pro, 2023.

[5] Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP journal on Image and Video Processing*, 2014: 1–28, 2014.

[6] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *ICML*, 2012.

[7] Bahri Batuhan Bilecen, Ahmet Berke Gokmen, and Aysegul Dundar. Dual encoder gan inversion for high-fidelity 3d head reconstruction from single images, 2024. URL https://arxiv.org/abs/2409.20530.

[8] Kevin W Bowyer, Karen Hollingsworth, and Patrick J Flynn. Image understanding for

iris biometrics: A survey. *Computer vision and image understanding*, 110(2):281–307, 2008.

[9] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, Anaheim, CA, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

[10] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. In *Foundation Models for Decision Making Workshop at Neural Information Processing Systems, 2022*, 2022. doi: 10.48550/arXiv.2210.04133.

[11] Aayush Kumar Chaudhary and Jeff B. Pelz. Privacy-preserving eye videos using rubber sheet model. In *ACM Symposium on Eye Tracking Research & Applications*, 2020.

[12] Yingshi Chen. GIRIST: Iris Recognition Software. https://github.com/closest-git/Girist, 2014. Accessed: [10/10/2023].

[13] Adam Czajka and Kevin W. Bowyer. Presentation attack detection for iris recognition: An assessment of the state-of-the-art. *ACM Computing Surveys (CSUR)*, 51(4):1–35, 2018. doi: 10.1145/3204947.

[14] John Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1167–1175, 2007. doi: 10.1109/TSMCB.2007.903540. URL https://doi.org/10.1109/TSMCB.2007.903540.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A

large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[16] James S. Doyle and Kevin W. Bowyer. Robust detection of textured contact lenses in iris recognition using bsif. *IEEE Access*, 3:1672–1683, 2015. doi: 10.1109/ACCESS.2015.2473111.

[17] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM CCS*, 2015.

[18] Abhishek Gangwar and Akanksha Joshi. Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. In *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305, 2016. doi: 10.1109/ICIP.2016.7532789. URL https://ieeexplore.ieee.org/document/7532789.

[19] Stephan J. Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S. Talathi. Openeds: Open eye dataset. *CoRR*, abs/1905.03702, 2019. URL http://arxiv.org/abs/1905.03702.

[20] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[21] Anil K. Jain, Arun Ross, and Umut Uludag. Biometric template security: Challenges and solutions. In *2005 13th European Signal Processing Conference*, pages 1–4, 2005.

[22] Javad Jarrahi. Iproov face biometrics deployed in ocbc atms in singapore pilot:

Biometric update, Mar 2021. URL https://www.biometricupdate.com/202103/iproov-face-biometrics-deployed-in-ocbc-atms-in-singapore-pilot.

[23] Brendan et al. John. Let it snow: Adding pixel noise to protect the user's identity. In *ACM Symposium on Eye Tracking Research & Applications*, 2020.

[24] Siamul Karim Khan, Patrick Tinsley, and Adam Czajka. Deformirisnet: An identity-preserving model of iris texture deformation, 2022. URL https://arxiv.org/abs/2207.08980.

[25] Oleg V Komogortsev, Sampath Jayarathna, Cecilia R Aragon, and Mechehoul Mahmoud. Biometric identification via an oculomotor plant mathematical model. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 57–60, 2010.

[26] Oleg V Komogortsev, Alexey Karpov, and Corey D Holland. Attack of mechanical replicas: Liveness detection with eye movements. *IEEE Transactions on Information Forensics and Security*, 10(4):716–725, 2015.

[27] Ketan Kotwal, Ibrahim Ulucan, Gökhan Özbulak, Janani Selliah, and Sébastien Marcel. Vrbiom: A new periocular dataset for biometric applications of hmd. *arXiv preprint arXiv:2407.02150*, 2024. URL https://arxiv.org/abs/2407.02150.

[28] Alexander Kupin, Benjamin Moeller, Yijun Jiang, Natasha Kholgade Banerjee, and Sean Banerjee. Task-driven biometric authentication of users in virtual reality (vr) environments. In *International Conference on Multimedia Modeling*, pages 55–67. Springer, 2019.

[29] Magic Leap. Magic leap 2: Iris id (beta). https://resources.magicleap.com/en-us/privacy/iris-unlock-id?locale=en-US, 2023.

[30] Young Won Lee and Kang Ryoung Park. Recent iris and ocular recognition methods in high- and low-resolution images: A survey. *Mathematics*, 10(12):2063, 2022. doi: 10.3390/math10122063. URL https://www.mdpi.com/2227-7390/10/12/2063.

[31] Lingzhi Li, Jianmin Bao, Ting Zhang, Dong Chen, Fang Wen, and Baining Guo. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4929, 2020. doi: 10.1109/CVPR42600.2020.00497. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_FaceShifter_Towards_High_Fidelity_and_Occlusion_Aware_Face_Swapping_CVPR_2020_paper.pdf.

[32] Dillon Lohr, Samantha Aziz, Lee Friedman, and Oleg V. Komogortsev. Gazebasevr, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality. *Scientific Data*, 10(1):177, 2023. doi: 10.1038/s41597-023-02117-5.

[33] Libor Masek and Peter Kovesi. MATLAB Source Code for a Biometric Identification System Based on Iris Patterns. The School of Computer Science and Software Engineering, The University of Western Australia, 2003.

[34] Anish Narkar and Brendan David-John. Swap it like its hot: Segmentation-based spoof attacks on eye-tracking images. In *ACM Symposium on Eye Tracking Research & Applications (ETRA)*. ACM, 2024.

[35] Kien Nguyen, Hugo Proença, and Fernando Alonso-Fernandez. Deep learning for iris recognition: A survey, 2022. URL https://arxiv.org/abs/2210.05866.

[36] Chinese Academy of Sciences' Institute of Automation (CASIA). Casia-iris-thousand, 2010. URL http://www.cbsr.ia.ac.cn/english/IrisDatabase.asp.

[37] Cristina Palmero, Abhishek Sharma, Karsten Behrendt, Kapil Krishnakumar, Oleg V.

Komogortsev, and Sachin S. Talathi. Openeds2020: Open eyes dataset, 2020. URL https://arxiv.org/abs/2005.03876.

[38] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Jian Jiang, Luis RP, Sheng Zhang, Pingyu Wu, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. URL https://arxiv.org/abs/2005.05535.

[39] Kiran B Raja, Ramachandra Raghavendra, and Christoph Busch. Video presentation attack detection in visible spectrum iris recognition using magnified phase information. *IEEE Transactions on Information Forensics and Security*, 10(10):2048–2056, 2015.

[40] Mehedi Hasan Raju, Dillon J. Lohr, and Oleg V. Komogortsev. Iris print attack detection using eye movement signals. In *Proceedings of the 2022 ACM Symposium on Eye Tracking Research and Applications*, pages 1–6, 2022. doi: 10.1145/3517031.3532521. URL https://doi.org/10.1145/3517031.3532521.

[41] Eduardo Ribeiro, Andreas Uhl, Fernando Alonso-Fernandez, and Reuben A. Farrugia. Exploring deep learning image super-resolution for iris recognition. page 2176–2180, August 2017. doi: 10.23919/eusipco.2017.8081595. URL http://dx.doi.org/10.23919/EUSIPCO.2017.8081595.

[42] Ioannis Rigas and Oleg V. Komogortsev. Eye movement-driven defense against iris print-attacks. *Pattern Recognition Letters*, 68:316–326, 2015. doi: 10.1016/j.patrec.2015.09.030. URL https://doi.org/10.1016/j.patrec.2015.09.030.

[43] 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 *40th Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, 2022.

[44] Ana F. Sequeira, Hélder P. Oliveira, João C. Monteiro, João P. Monteiro, and Jaime S. Cardoso. Mobilive 2014 - mobile iris liveness detection competition. In *IEEE International Joint Conference on Biometrics*, pages 1–6. IEEE, 2014. doi: 10.1109/BTAS.2014.6996290.

[45] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[46] Ars Staff. Microsoft unveils hololens 2: Twice the field of view, eye tracking. https://arstechnica.com/gadgets/2019/02/microsoft-unveils-hololens-2-twice-the-field-of-view-eye-tracking/, 2019.

[47] Patrick Tinsley, Adam Czajka, and Patrick Flynn. Haven't i seen you before? assessing identity leakage in synthetic irises, 2022.

[48] Inmaculada Tomeo-Reyes, Arun Ross, Antwan Clark, and Vinod Chandran. A biomechanical approach to iris normalization. In *Proceedings of the 8th IAPR International Conference on Biometrics (ICB)*, pages 336–343, 2015. doi: 10.1109/ICB.2015.7139095. URL https://www.cse.msu.edu/~rossarun/pubs/TomeoReyesIrisNormalization_ICB2015.pdf.

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[50] Sushma Krupa Venkatesh, Raghavendra Ramachandra, K Bommanna Raja, and Christoph Busch. A new multi-spectral iris acquisition sensor for biometric verification and presentation attack detection. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 47–54. IEEE, 2019.

[51] Zhenan Wei, Tieniu Tan, and Zhenan Sun. Nonlinear iris deformation correction based on gaussian model. In *Advances in Biometrics. ICB 2007. Lecture Notes in Computer Science*, volume 4642, pages 780–789. Springer, 2007. doi: 10.1007/978-3-540-74549-5_82. URL https://link.springer.com/chapter/10.1007/978-3-540-74549-5_82.

[52] Zhengyang Wu, Srivignesh Rajendran, Tarrence van As, Joelle Zimmermann, Vijay Badrinarayanan, and Andrew Rabinovich. Magiceyes: A large scale eye gaze estimation dataset for mixed reality, 2020. URL https://arxiv.org/abs/2003.08806.

[53] Shivangi Yadav and Arun Ross. iwarpgan: Disentangling identity and style to generate synthetic iris images. *arXiv preprint arXiv:2305.12596*, 2023. URL https://arxiv.org/abs/2305.12596.

[54] Shivangi Yadav, Cunjian Chen, and Arun Ross. Synthesizing iris images using rasgan with application in presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. doi: 10.1109/CVPRW.2019.00010. URL https://openaccess.thecvf.com/content_CVPRW_2019/papers/Biometrics/Yadav_Synthesizing_Iris_Images_Using_RaSGAN_With_Application_in_Presentation_Attack_CVPRW_2019_paper.pdf.

[55] Zijie Zhang, Changhong Fu, Yongkang Cao, Mengyuan Li, and Haobo Zuo. Livedet: Lightweight density-guided adaptive transformer for online on-device vessel detection. *IEEE Robotics and Automation Letters*, 10(6):5513–5520, 2025. doi: 10.1109/LRA.2025.3559834.