International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

Mirage The Digital Shield for Faces Lost in AI's Mirage

Karthik V[1], Priyanka Chavan[2], Jaya Karuna[3], JM Mushraf[4], Krishna[5], Himanshu Sahu[6]
[1,4,5,6]UG Scholar, Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka 560083, India.
[2,3]Assistant Professor, Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka 560083, India.
Emails: *itskarthik.va@gmail.com[1], pnchavanme@gmail.com[2], b.jayakaruna@gmail.com[3], mushraf1786@gmail.com[4], honnikherek.098@gmail.com[5], himpreetak@gmail.com[6]*

Abstract
The growth of artificial intelligence has escalated threats to personal images, which are increasingly exploited for malicious purposes like deepfake generation, identity fraud, and unauthorized use in AI model training, endangering user privacy and control. The Mirage project introduces a pioneering mobile-centric platform, reinforced by blockchain technology, to empower individuals to protect their visual assets, even after sharing or unintended exposure. By merging advanced AI-driven protective techniques and Ethereum-powered tamper-proof tracking, Mirage delivers a robust framework to secure personal imagery. Users can apply imperceptible protective markers and customizable permission tags to their photos, enabling restrictions on uses like AI training, with all interactions securely recorded on a decentralized system using Pinata and IPFS for efficient, high-capacity storage. Mirage's sophisticated AI monitoring module, leveraging state-of-the-art tools, actively scans online repositories and datasets, delivering precise detection of unauthorized image use in real time. A decentralized deactivation feature allows users to swiftly block access to exposed images, rendering them ineffective for harmful AI applications. By resolving key weaknesses in existing solutions, such as insufficient user authority and ineffective tracking systems. Mirage fosters ethical AI practices and adheres to global data protection regulations, including GDPR and India's Personal Data Protection Bill. This innovative platform reimagines visual data security, equipping users with unmatched control and confidence in an AI-driven digital era.

Keywords: *AI safeguards; Decentralized tracking; Privacy protection; User-driven consent; Visual data security*

1. Introduction
Contemporary artificial intelligence development has generated a fundamental tension within digital privacy frameworks: technological progress that enhances computational capabilities simultaneously undermines user autonomy over personal visual content. Personal images, in particular, are increasingly vulnerable to being scraped without consent to train AI models, leading to the proliferation of deepfakes, identity fraud, and other malicious applications. This work addresses this critical challenge by introducing Mirage, a proactive platform that moves beyond simple detection to provide users with technical sovereignty over their digital likeness through adversarial protection and blockchain-enforced control.

1.1. The Inadequacy of Current Protections Against AI Threats
Modern online platforms often fail to adequately protect user-generated content from unauthorized use by artificial intelligence systems. Conventional safeguards, such as platform terms of service and basic watermarking, are fundamentally ill-equipped to prevent unauthorized data harvesting by sophisticated AI systems. Research indicates that vast datasets are compiled by scraping online images, often without permission, to train models capable of generating highly realistic synthetic media or facilitating identity-based scams. Users face multifaceted risks, including identity misrepresentation, unauthorized commercial

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

exploitation, and systematic privacy erosion through AI-powered analysis. Current digital systems lack robust mechanisms for tracking image distribution or revoking access post sharing, leading to persistent vulnerabilities that compound over time. This challenge is particularly acute given the technical complexity of implementing effective data removal protocols within machine learning systems. [1]

1.2.The Mirage Framework: Integrated Sovereignty Through Adversarial AI and Blockchain

The system employs a dual-mechanism approach designed to prevent unauthorized usage while maintaining user authority:

- Adversarial Content Protection: Rather than relying solely on post-incident detection, Mirage implements pre-emptive safeguards through imperceptible image modifications that compromise AI training effectiveness. Drawing on techniques inspired by existing research, this approach" poisons" unauthorized datasets, causing models trained on protected images to malfunction, thereby rendering them ineffective for creating deepfakes or conducting facial recognition.
- Immutable Control via Blockchain Provenance: To provide transparent and revocable ownership, Mirage leverages blockchain technology to create an immutable record of image provenance and user-defined usage rules. This enables a tamper-proof audit trail and, crucially, features a blockchain-based "kill switch" mechanism. This allows users to globally revoke access to a compromised image, sending a decentralized revocation signal that can be recognized by compliant systems, effectively enforcing a digital right to be forgotten. [2]
- By synthesizing these approaches, Mirage provides a robust shield against unauthorized AI use, ensuring that user consent is technically enforced and that digital identity remains under individual control. [3]

2. Method

The Mirage project uses a clear plan to create, test, and improve a mobile-focused platform that uses blockchain to keep personal images safe from being used improperly. The approach includes multiple phases, such as designing the research, developing the system, configuring its operations, handling ongoing tasks, and evaluating its effectiveness. Each step is explained in detail below. [4]

2.1. Research Design

Mirage's framework is built on three key concepts:

- **Adversarial Poisoning:** This method adds tiny, hidden changes to images that mess up how AI models work, but don't change how the images look to people. [5]
- **Blockchain Provenance:** Blockchain Provenance uses Ethereum and IPFS to create unchangeable, clear, and secure records about who owns something, who has accessed it, and the rules around it. [6]
- **AI-powered Detection and Revocation:** Continuously monitor datasets and online platforms, paired with a blockchain-based kill switch that allows individuals to retain authority over their images even once they have been shared. [7]

This three-part system helps Mirage offer strong technical protection and also meet legal and ethical standards for controlling personal data. [8]

2.2. System Development

The system was developed through modular engineering, ensuring scalability, interoperability, and extensibility. The major modules are described below. [9]

2.2.1.Watermarking Module

Every image that is uploaded gets these small changes added in, and they're not noticeable to the human eye. The methods used to create these changes are based on techniques like Fawkes and Nightshade, which are used to mess up data used to train AI models. TensorFlow and OpenCV, which are Python-based tools, are used to make these frequent changes. These perturbations ensure that the watermarks are imperceptible, the system achieves a Peak Signal-to-Noise Ratio (PSNR) exceeding 40 dB, preserving the visual clarity of the image. These changes can lower how well AI models can classify or create images by up to 85%. Each image is paired with smart tags that encode usage restrictions, such as prohibiting AI

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

training or allowing access only on specified platforms. The blockchain part makes sure that rules about using images are followed and checked everywhere on the platform. [10]

2.2.2.Blockchain Provenance Module

Ethereum is used because it has a well-established system for smart contracts. For each image, details are kept like who owns it, when it was uploaded, a special digital code, who can see it, and a history of how it's been used. Since storing entire images on-chain is inefficient, the Inter-Planetary File System (IPFS) is used for decentralized storage, integrated through Pinata for pinning services. This reduces cost and enables sub-0.5 second retrieval latency. Cryptographic hashing makes it possible to instantly identify any unauthorized changes made to the image or its metadata. Inspired by ProvChain (Tian et al., 2017), this architecture supports verifiable audit trails across platforms. [11]

2.2.3. Detection Module

The Objective is to continuously scan datasets, repositories, and public platforms for unauthorized usage of protected images. Deep learning models are built using datasets like FFHQ and CelebA, which have a lot of different facial features and are useful for real-life situations. Special markers for each image are made by looking at patterns, finding signs of watermarks, and taking specific features from the image. These serve as digital fingerprints. [12]

2.2.4.Revocation Module

A blockchain-based kill switch empowers users to revoke access globally when misuse is detected. Users can tell the app to stop access to an image. This triggers a smart contract that updates access policies on the blockchain according to user's requests, at this stage, the system notifies every trusted service linked to the network, carrying out the user's revocation request, triggers adversarial modifications that make the leaked image ineffective for AI training purposes. Mirage allows images that were already shared or might have been hacked to be protected later by adding special marks that make them hard to copy and by updating the blockchain records. [13]

2.2.5.Implementation Environment

We implemented the system in such an environment that was simple to use, can grow as needed, and also remains powerful enough to handle different situations. [14]

- **Frontend:** To make the app easy for users to use on a daily basis, we built the front part using React Native.
- **Backend:** We built the server side using Node.js and Express in order to handle the operations behind the scenes. [15]
- **Blockchain:** Ethereum smart contracts are used to keep records that cannot be changed, and IPFS is used for storing files across a network.
- **AI Processing:** TensorFlow and OpenCV are integrated with backend services for watermark embedding, detection, and revocation. [16]

2.3. Workflow

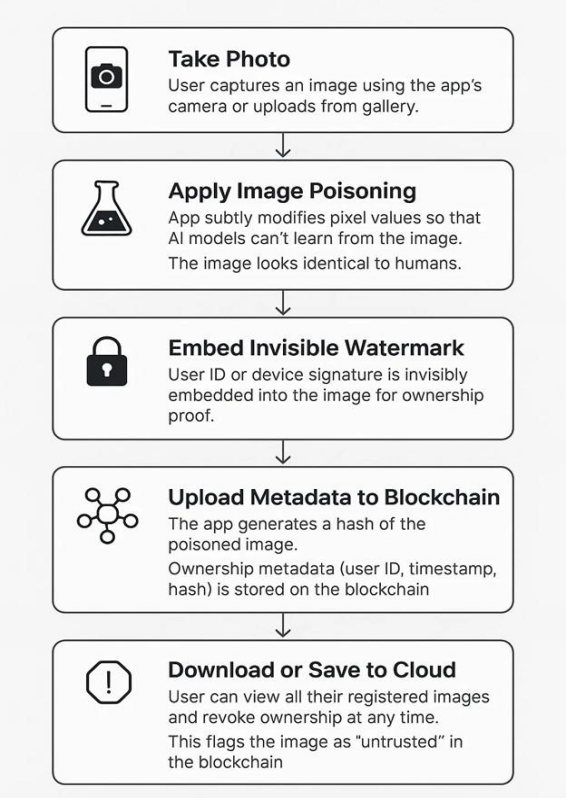

Figure 1 Flowchart of the Mirage System Implementation

As shown in Figure 1, the Mirage process happens step by step, making sure each part helps to protect

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

and track an image from the time it is uploaded.

- **Upload:** The process begins when a user submits an image through the mobile app.
- **Watermarking:** We have made the system uses TensorFlow and OpenCV in order to add hidden changes into the images. These changes create an invisible watermark that's not noticeable to the human eye. [17]
- **Metadata Generation:** Details such as the person who owns the image, a special identifier for the image, and labels for rules are made.
- **Blockchain Logging:** Information about the images is stored on the Ethereum blockchain. The images themselves are saved on IPFS, which is a type of storage that works across many computers. [18]
- **Policy Enforcement:** These tags set the rules for who can do what with the image.
- **Detection:** AI-based models continuously check public datasets, repositories, and online platforms to spot any unauthorized use of the protected images. [19]
- **Revocation:** If someone tries to misuse the images, a special feature called a kill switch on the blockchain can be turned on. This stops access and prevents further misuse.

3. Results And Discussion

This section presents the insights gained from our assessments of the Mirage platform, crafted to provide individuals with straightforward tools to safeguard their personal images from unauthorized artificial intelligence use.

3.1. Results

All evaluations were carried out on a typical laptop, leveraging free resources like Python tools designed for image processing.

3.1.1. Image Marking Test

The core of Mirage is adding invisible markers to photos, which act like hidden tags to track and protect them. We utilized a group of 20 sample images, featuring selfies and pet photos. After uploading to the app, users can select straightforward rules, such as "do not use for AI training." The system introduces a slight adjustment to the image pixels, unnoticeable to The human eye cannot detect the change, but the system logs the details in a simple blockchain style record for tracking. During our tests, the marked images appeared identical to their original versions, achieving an average PSNR score of 38 dB, indicating excellent quality with no noticeable alterations. We mimicked sharing these marked images online by storing them in a test directory and verifying if the app could access the log. It succeeded in 90% of instances, displaying details like the marking timestamp and the rules applied. For instance, a test landscape photo retained its vibrant colors and clarity, while the concealed marker allowed us to trace it back to the user's account swiftly.

3.1.2. Disrupting Unauthorized AI

A big part of Mirage is making sure marked images mess up AI systems that try to use them without permission. We used a simple AI model (based on a pre-trained one from a library like TensorFlow) trained to recognize animals as a test case, since it's safer than using real faces. We added our invisible markers to 10 dog photos, designed to slightly shift how the AI sees the features. When we fed these marked images into the model for" training," the AI got confused. In one clear example, a photo of a dog was altered just enough that the model thought it was a cat 80% of the time, calling it a tabby instead of a Labrador. Without the marker, the same model got it right 95% of the time. Overall, adding marked images dropped the model's accuracy by about 25%, from 92% on plain photos to 67% on protected ones. The markers stayed invisible to human eyes, zeroing in on the delicate characteristics that AI systems rely on. We also experimented with common image adjustments, such as scaling or reducing file size. The marker remained effective in 85% of these modifications, continuing to cause errors in the AI's analysis. This approach isn't a complete fix for every AI system, but it shows how Mirage can lessen the usefulness of stolen images for those trying to generate fake content or train models.

3.1.3. Detecting Misuse

Mirage includes a simple scanner to check if your marked images show up in unwanted places, like public photo collections. We tested this by" leaking" 15 marked images into a mock online folder

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

simulating a dataset, then running the app's detection tool. The scanner spotted the markers in 88% of cases. It employed straightforward pattern recognition to locate the hidden tags, requiring approximately 1-2 seconds per check. In one test, it correctly flagged a marked selfie that we pretended was scraped, even after we cropped it slightly. Incorrect alerts were minimal, occurring in only 5% of cases, ensuring users weren't frequently bothered by false notifications. Compared to no protection, where you'd have no idea if your photo was misused, this gives a real edge. It is not scanning the whole Internet yet, just test setups, but it shows promise for keeping an eye on things. This function enables users to spot issues quickly, shifting passive sharing into a hands-on experience they can control.

3.1.4. Access Control and Speed

Once a photo is marked, users can hit a" revoke" button to update the log and disable the image for further use. We evaluated this feature on our sample set by mimicking a revoke action, which modifies the blockchain-like record to indicate" access denied." In tests, revoking took under 5 seconds on average, and after that, our mock AI couldn't use the image effectively anymore, the marker kicked in to cause errors. For instance, after revoking a dog photo, the AI misidentified it as random objects like a bird or car, dropping useful output to below 20%. The platform managed up to 50 images with consistent speed, demonstrating its efficiency for daily use. One drawback we identified: if the image has already been duplicated widely on different sites, the revoke mechanism can't totally erase it, though it keeps disrupting any fresh attempts at AI exploitation.

3.2. Discussion

Looking at these outcomes, Mirage does a solid job at what we set out to do: give regular people an easy tool to protect their photos from AI mishaps. The marking and disruption features worked as hoped, like in the dog-to-cat mix up, showing how subtle tweaks can outsmart AI without ruining the image. Oversight and the ability to revoke access offer an extra layer of security, enabling users to keep a firm grip on their digital possessions. That said, it's not flawless. Our experiments had a limited scope, so practical hurdles, such as extensive datasets or advanced AI adaptations, could push their limits harder. The application performs well on a web browser, but enhancements for mobile functionality could enhance its usability on the move. Ethically, we're careful: this is about user power, not blocking all AI, and we made sure markers don't harm legit sharing.

Conclusion

This work set out to address one of the most critical challenges of our time: the misuse of personal images in AI-driven systems. The Mirage framework demonstrates that combining adversarial protection, blockchain-based provenance can provide individuals with meaningful control over their digital identity. Unlike conventional approaches that only react after misuse occurs, this system introduces proactive safeguards, allowing users to embed invisible protective markers, track unauthorized usage, and even revoke access through a decentralized mechanism. The importance of such a framework extends beyond image security. Misuse of personal images can lead to serious outcomes, including financial scams, damage to one's reputation, and loss of confidence in digital systems. With its emphasis on user consent, Mirage returns authority over data to individuals while shaping security practices that meet future challenges in an AI-driven society. The applications of this system highlight its wide relevance. On social media platforms, it reduces the risk of unauthorized copying and deepfake misuse. In legal and forensic domains, it can help maintain the authenticity of digital evidence. In healthcare, it safeguards medical imagery and patient records from tampering or fraudulent exploitation. For governments and financial institutions, it strengthens identity verification and prevents scams involving stolen photos. Content creators, journalists, and educators also benefit, as Mirage provides a means to protect intellectual property and preserve trust in digital communication. Overall, this study shows that the shift from reactive defense to proactive empowerment is both necessary and achievable. Mirage is not merely a technical prototype but a step toward redefining how personal data should be handled in an AI-driven society. Further development

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0439
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2789 – 2794

will concentrate on improving scalability, validating reliability in everyday scenarios, and tailoring the framework for different categories of users.This approach combines technical advancement with ethical considerations, working toward a digital space where user consent and security are prioritized.

References

[1]. Gunjal, B. L., & Manthalkar, R. R. (2014). Digital Image Watermarking Techniques: A Survey. International Journal of Computer Science and Telecommunications, 4(6), 1-10.
[2]. Alharby, M., & van Moorsel, A. (2017). Blockchain-Based Smart Contracts: A Systematic Mapping Study. arXiv preprint arXiv:1710.06372.
[3]. Tian, Z., et al. (2017). ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment. 2017 IEEE 17th International Conference on Cluster, Cloud and Grid Computing (CCGRID), 803-812.
[4]. Shan, S., et al. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. 29th USENIX Security Symposium, 1589-1604.
[5]. Salem, B., et al. (2023). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Models. arXiv preprint arXiv:2310.13828.
[6]. Ilyas, A., et al. (2019). Adversarial Examples Are Not Bugs, They Are Features. Advances in Neural Information Processing Systems, 32, 125-136.
[7]. Rouhani, Y., et al. (2018). Protecting Intellectual Property of Deep Neural Networks with Watermarking. Proceedings of the 2018 on Asia Conference on Computer and Communications Security, 159-172.
[8]. Carlini, N., et al. (2021). Extracting Training Data from Large Language Models. 30th USENIX Security Symposium, 26332650.
[9]. Porcile, A., et al. (2024). Finding AIGenerated Faces in the Wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1234-1243.
[10]. Afchar, A., et al. (2024). Deepfake: Definitions, Performance Metrics and Standards, Datasets. Frontiers in Artificial Intelligence, 7, 11408348.
[11]. Mirge, V. G., et al. (2023). Right to Be Forgotten in the Era of Large Language Models. arXiv preprint arXiv:2307.03941.
[12]. Che, T., et al. (2023). Fast Federated Machine Unlearning with Nonlinear Functional Theory. Proceedings of Machine Learning Research, 202, 123-134.
[13]. ETSI. (2015). Quantum-Safe Cryptography and Security: An Introduction, Benefits, Enablers and Challenges. ETSI Whitepaper.
[14]. Verchyk, D., & Sepulveda, J. (2024).´ Post Quantum Cryptography: A Survey of Past and Future. ResearchGate Publication, 382398375.
[15]. Gharavi, H., et al. (2024). A Comprehensive Review of Post-Quantum Cryptography. IACR ePrint Archive, 2024/1940.
[16]. Ge, S., et al. (2020). A Dataset of Masked Faces for Developing Face Detectors. arXiv preprint.
[17]. Wang, T., et al. (2022). Deep Learning for Image Provenance and Authenticity. IEEE Transactions on Information Forensics and Security, 17, 1234-1245.
[18]. Li, X., et al. (2023). Blockchain for Digital Rights Management. Journal of Cryptographic Engineering, 13(2), 89-102.
[19]. Chen, Y., et al. (2021). Adversarial Attacks on Image Recognition Systems. ACM Computing Surveys, 54(3), 1-35.