AI and Cybersecurity

DATA POISONING WON'T SAVE YOU FROM FACIAL RECOGNITION

Written by: Evani Radiya-Dixit, Nicholas Carlini, Sanghyun Hong and Florian Tramèr.

Presented by: **Mashal Zainab** on 18th December, 2023.
Code Repository:
https://github.com/mashalbhatti/AI_and_Cybersecurity_Project_FaceCure

Contents of this Presentation

DATA POISONING WON'T SAVE YOU FROM FACIAL RECOGNITION

Evani Radiya-Dixit
Stanford University

Sanghyun Hong
Oregon State University

Nicholas Carlini
Google

Florian Tramèr
Stanford University, Google

ABSTRACT

Data poisoning has been proposed as a compelling defense against facial recognition models trained on Web-scraped pictures. Users can perturb images they post online, so that models will misclassify future (unperturbed) pictures.

We demonstrate that this strategy provides a false sense of security, as it ignores an inherent asymmetry between the parties: users' pictures are perturbed *once and for all* before being published (at which point they are scraped) and must thereafter fool *all future models*—including models trained adaptively against the users' past attacks, or models that use technologies discovered after the attack.

We evaluate two systems for poisoning attacks against large-scale facial recognition, *Fawkes* (500,000+ downloads) and *LowKey*. We demonstrate how an "oblivious" model trainer can simply *wait* for future developments in computer vision to nullify the protection of pictures collected in the past. We further show that an adversary with black-box access to the attack can (i) train a robust model that resists the perturbations of collected pictures and (ii) detect poisoned pictures uploaded online.

We caution that facial recognition poisoning will not admit an "arms race" between attackers and defenders. Once perturbed pictures are scraped, the attack cannot be changed so any *future* successful defense irrevocably undermines users' privacy.

1 INTRODUCTION

Facial recognition systems pose a serious threat to individual privacy. Various companies routinely scrape the Web for users' pictures to train large-scale facial recognition systems (Hill, 2020a; Harwell, 2021), and then make these systems available to law enforcement agencies (Lipton, 2020) or private individuals (Harwell, 2021; Mozur & Krolik, 2019; Wong, 2019).

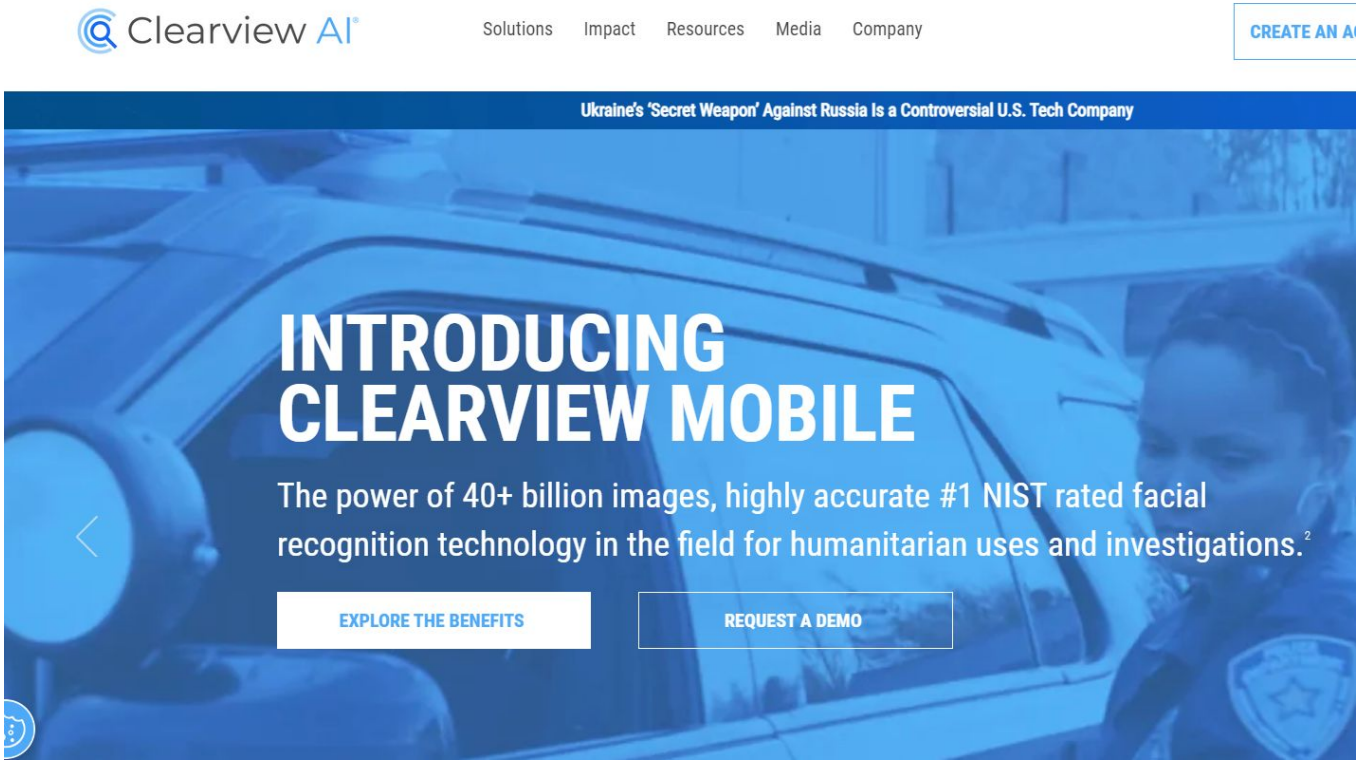
A growing body of work develops tools to allow users to fight back, using techniques from *adversarial machine learning* (Sharif et al., 2016; Oh et al., 2017; Thys et al., 2019; Kulynych et al., 2020; Shan et al., 2020; Evtimov et al., 2020; Gao et al., 2020; Xu et al., 2020; Yang et al., 2020; Komkov & Petiushko, 2021; Cherepanova et al., 2021a; Rajabi et al., 2021; Browne et al., 2020).

One approach taken by these tools lets users perturb any picture before they post it online, so that facial recognition models that train on these pictures will become *poisoned*. The objective is that when an *unperturbed* image is fed into the poisoned model (e.g., a photo taken by a stalker, a security camera, or the police), the model misidentifies the user. This approach was popularized by *Fawkes* (Shan et al., 2020), an academic image-poisoning system with 500,000+ downloads and covered by the New York Times (Hill, 2020b), that promises "strong protection against unauthorized [facial recognition] models". Following Fawkes' success, similar systems have been proposed by academic (Cherepanova et al., 2021a; Evtimov et al., 2020) and commercial (Vincent, 2021) parties.

This paper shows that these systems (and, in fact, any poisoning strategy) **cannot protect users' privacy**. Worse, we argue that these systems offer a false sense of security. There exists a class of privacy-conscious users who might have otherwise never uploaded their photos to the internet; however who now might do so, under the false belief that data poisoning will protect their privacy. These users are now *less private* than they were before. Figure 1 shows an overview of our results.

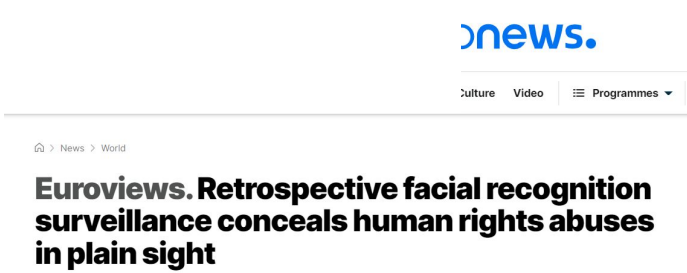
Part 1: Paper Explanation

Facial Recognition for Mass Surveillance

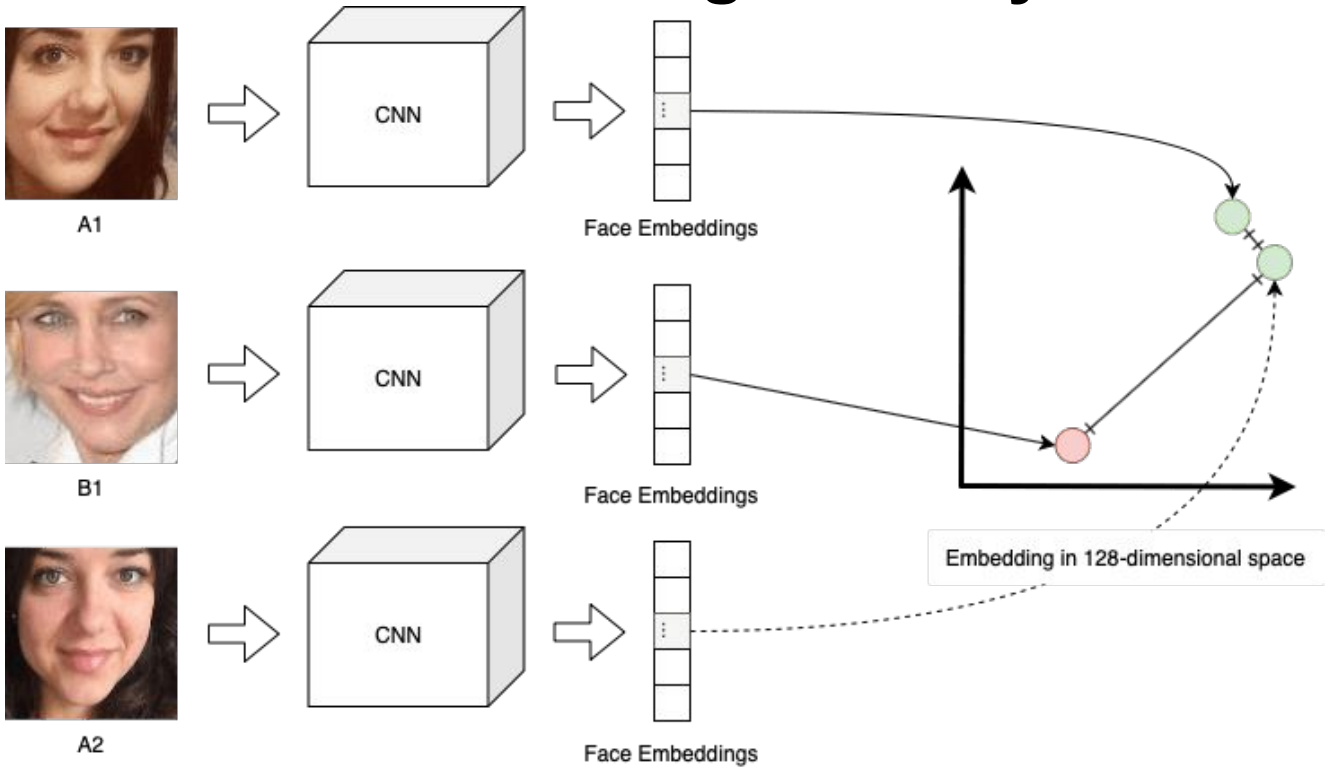


Source: Clearview AI (n.d.). *Clearview AI | The World's Largest Facial Network*. [online] Clearview AI. Available at: https://www.clearview.ai/ .

Facial Recognition for Mass Surveillance

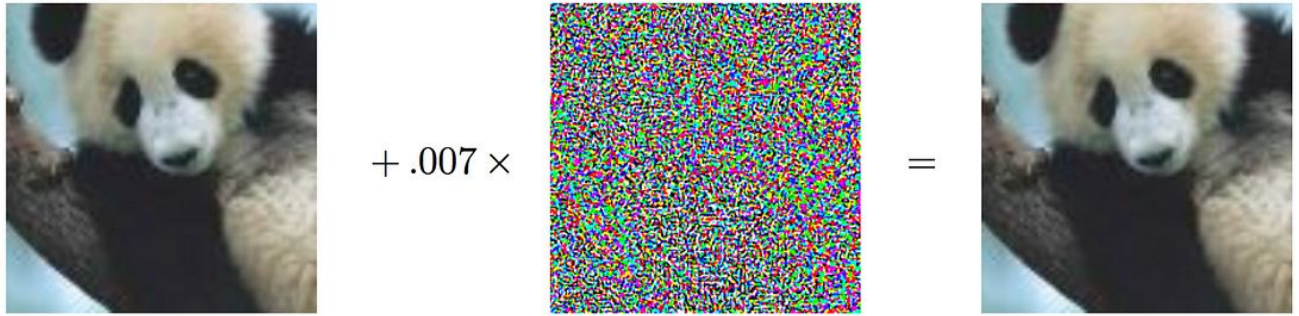


How does a Facial Recognition System work?



Jaiswal, A. (2022). *Face Recognition System Using Python*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/04/face-recognition-system-using-python/.

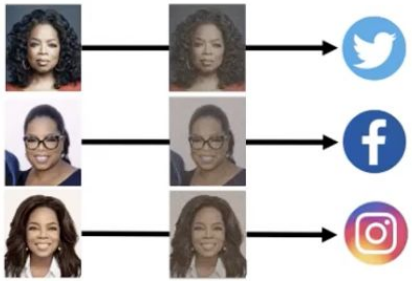
Data Poisoning for Facial Recognition - Adversarial Machine Learning



Source: guest_blog (2022). *Machine Learning: Adversarial Attacks and Defense*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/09/machine-learning-adversarial-attacks-and-defense/.

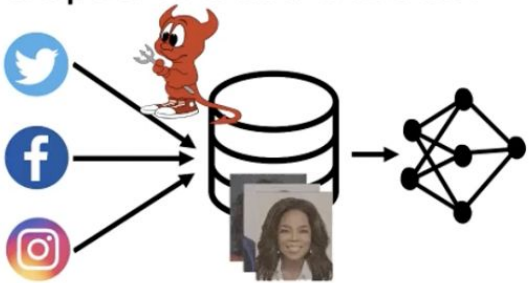
Data Poisoning for Facial Recognition - Adversarial Machine Learning

Users perturb pictures they post online

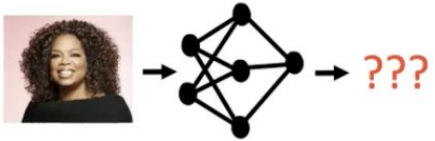


Users' friends can still recognize the pictures

Online pictures are scraped to build a model



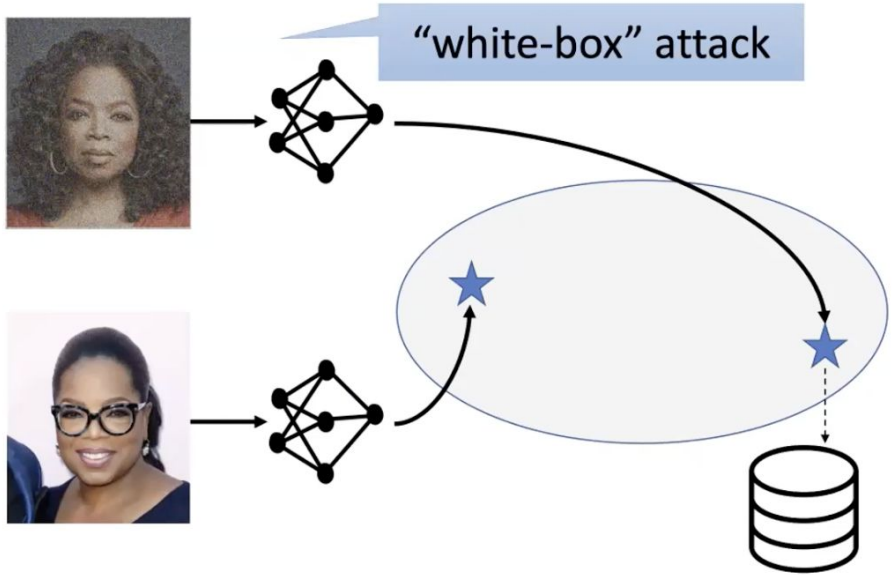
Unperturbed test pictures aren't recognized



Unperturbed picture taken by the police, or a stalker, etc.

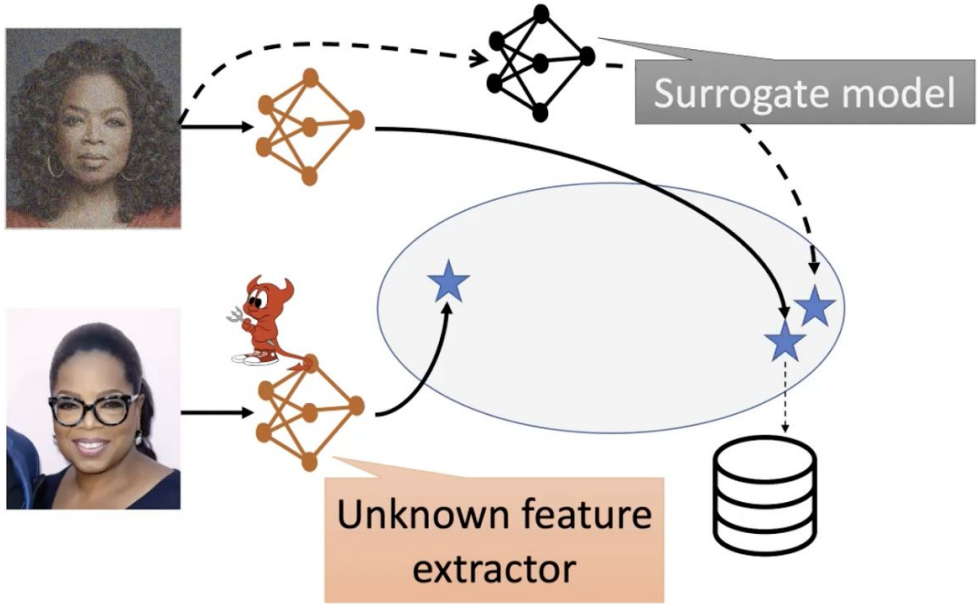
Data Poisoning for Facial Recognition - Adversarial Machine Learning

Poisoning is easy if the extractor is *fixed & known*.



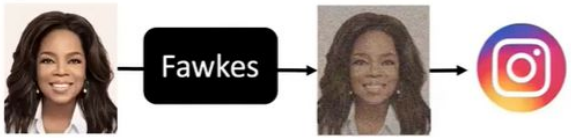
Data Poisoning for Facial Recognition - Adversarial Machine Learning

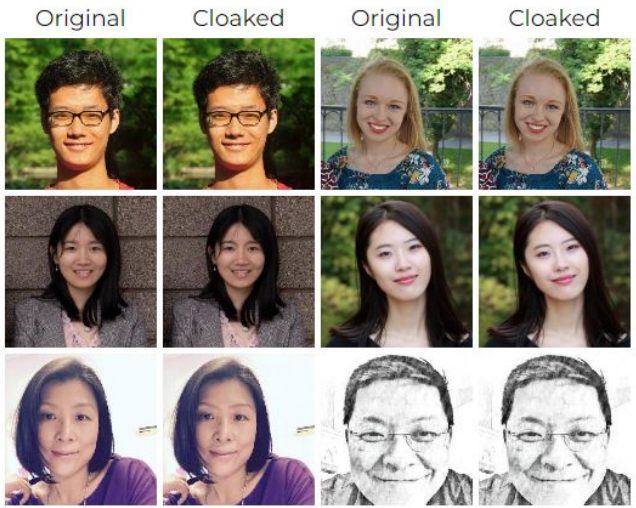
The attack should transfer to unknown extractors.



Fawkes - Image "Cloaking" for Personal Privacy

- Protecting Privacy against Unauthorized Deep Learning Models.
- Academic Image Poisoning system.
- Developed by the SAND Lab at University of Chicago.
- Published in July 2020.
- More than 500,000 downloads in the same year.
- Promises “strong protection against unauthorized [facial recognition] models”

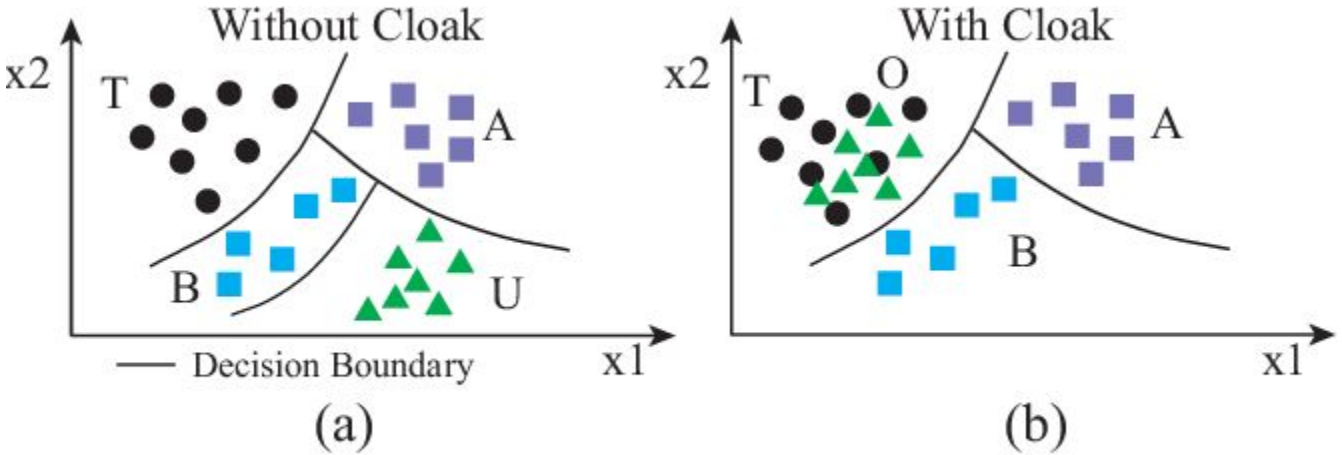




Source: Uchicago.edu. (2020). *Fawkes*. [online]

Available at: https://sandlab.cs.uchicago.edu/fawkes/.

How Fawkes Image Cloaking Works?



Source: Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H. and Zhao, B. (2020). *Open access to the Proceedings of the 29th USENIX Security Symposium is sponsored by USENIX. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models Fawkes: Protecting Privacy against Unauthorized Deep Learning Models*. [online] Available at: https://www.shawnshan.com/files/publication/fawkes.pdf [Accessed 17 Dec. 2023].

Fawkes v0.3 to Fawkes v1.0

News: Jan 28, 2021. It has recently come to our attention that there was a significant change made to the Microsoft Azure facial recognition platform in their backend model. Along with general improvements, our experiments seem to indicate that Azure has been trained to lower the efficacy of the *specific version* of Fawkes that has been released in the wild. We are unclear as to why this was done (since Microsoft, to the best of our knowledge, does not build unauthorized models from public facial images), nor have we received any communication from Microsoft on this. However, we feel it is important for our users to know of this development. We have made a major update (v1.0) to the tool to circumvent this change (and others like it). Please download the newest version of Fawkes below.

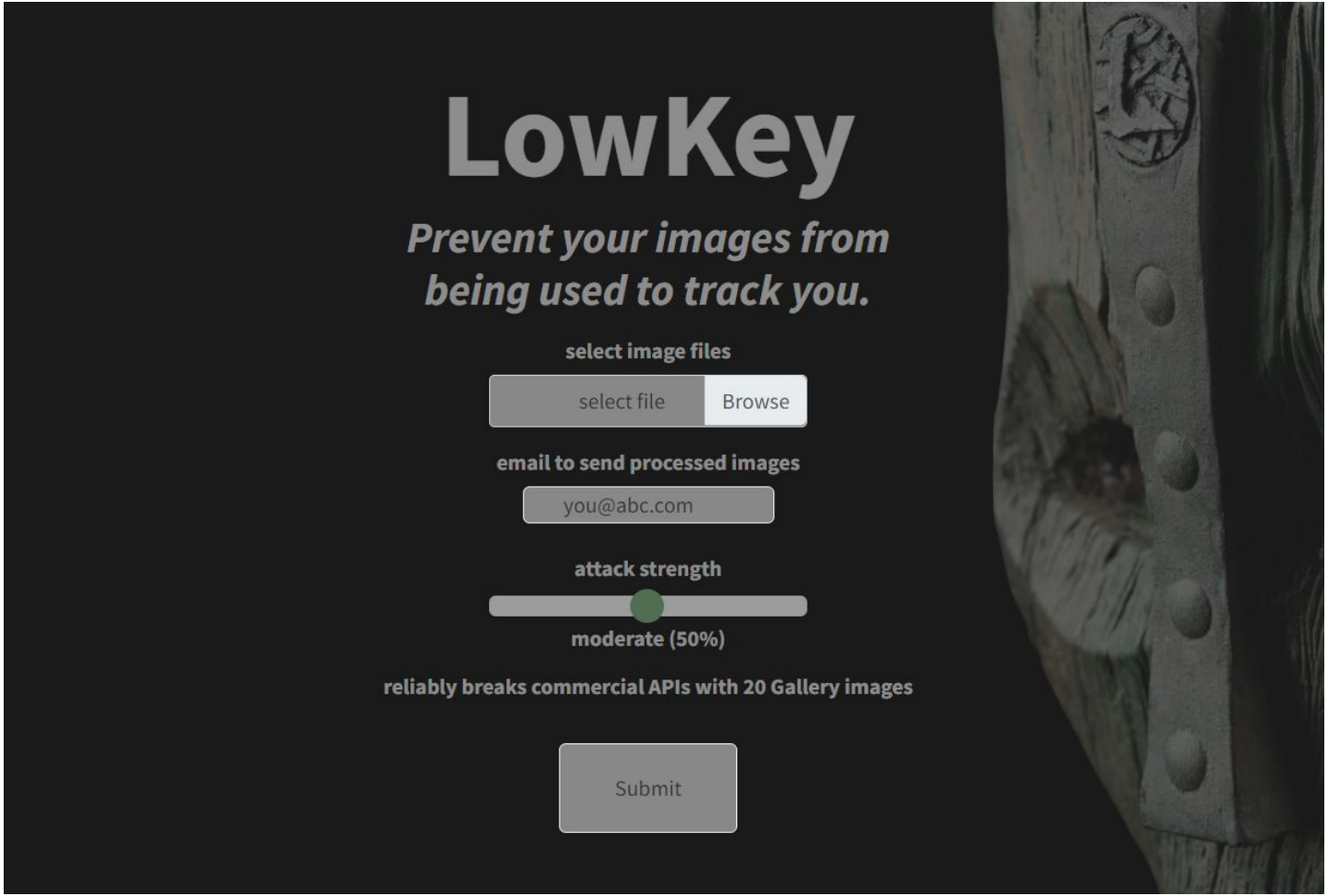
Source: Uchicago.edu. (2020). *Fawkes*. [online] Available at: https://sandlab.cs.uchicago.edu/fawkes/.

Lowkey

- Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition.
- Developed by Researchers University of Maryland and US Naval Academy.
- Published in January 2021.

Figure 1: Top: original images, Bottom: protected by LowKey.

Source: Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G. and Goldstein, T. (n.d.). *LOWKEY: LEVERAGING ADVERSARIAL ATTACKS TO PROTECT SOCIAL MEDIA USERS FROM FACIAL RECOGNITION*. [online] Available at: https://openreview.net/pdf?id=hJmtwocEqzc [Accessed 12 Dec. 2023].



Source: lowkey.umiacs.umd.edu. (n.d.). *Lowkey*. [online] Available at: https://lowkey.umiacs.umd.edu/ [Accessed 12 Dec. 2023].

This paper focuses on

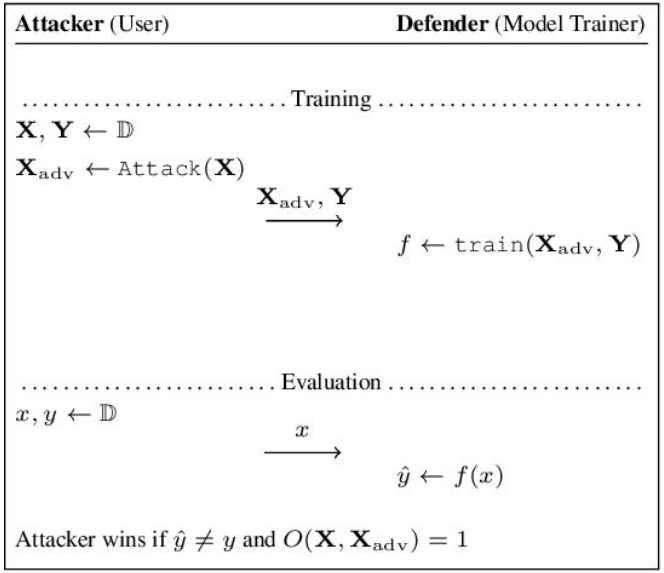
- The claim that these systems cannot protect users' privacy.
- Or any other poisoning strategy.
- These systems provide users with a fake sense of security.
- And they are less secure than before.
- Poisoning against facial recognition will not lead to an "arms race".
- Where new attacks can continuously counteract new defences.

X (formerly Twitter). (n.d.). *https://twitter.com/florian_tramer/status/1409882206412156932*. [online] Available at: https://twitter.com/florian_tramer/status/ 1409882206412156932 [Accessed 12 Dec. 2023].

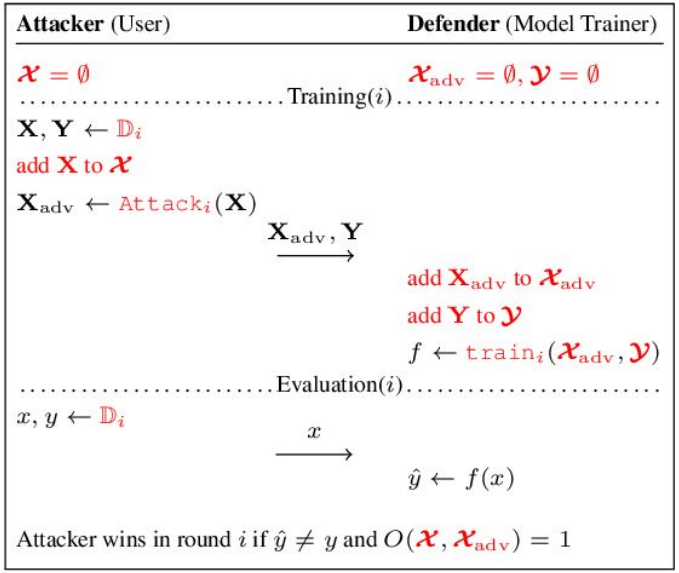
And shows that

- An adaptive model trainer with black-box access to Fawkes and LowKey can train a robust model that resists poisoning attacks and correctly identifies all users with high accuracy.
- An adaptive model trainer can also detect perturbed pictures with near-perfect accuracy.
- Fawkes and LowKey are already broken by newer facial recognition models that appeared less than a year after the attacks were introduced.
- Achieving robustness against poisoning attacks need not come at a cost in clean accuracy (in contrast to existing defenses against adversarial examples).

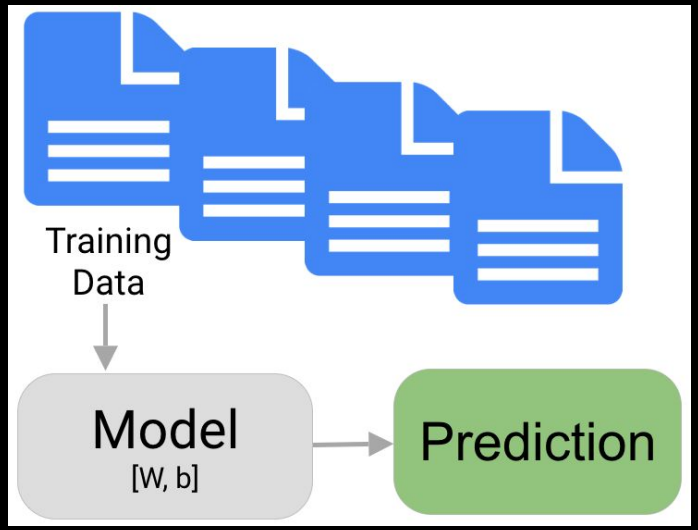
Poisoning Attack Games



(a) **Game 1: Static game.** The attacker creates a clean-labeled poisoned training set $(\mathbf{X}_{\text{adv}}, \mathbf{Y})$ and the defender trains a model f, which is evaluated on *unperturbed* inputs x. The attacker wins if f misclassifies x and the poisoned data $\mathbf{X}_{\text{adv}}$ is "close" to the original data $\mathbf{X}$ (according to an oracle O).



(b) **Game 2: Dynamic game.** In each round $i \geq 1$, the attacker sends new poisoned data to the defender. The defender may train on *all* the training data $(\boldsymbol{\mathcal{X}}_{\text{adv}}, \boldsymbol{\mathcal{Y}})$ it collected over prior rounds. The strategies of the attacker and defender may change between rounds.



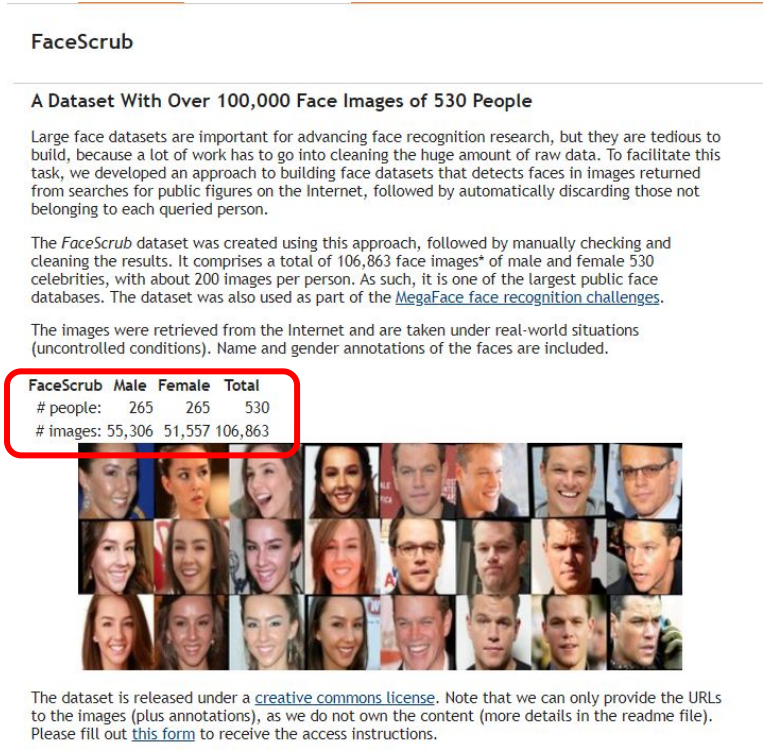
Part 2: Reproduction of the Experiments

Dataset and Evaluation Metric

The experiments in the paper are performed with the FaceScrub dataset (Ng & Winkler, 2014), which contains over **50,000 images of 530 celebrities**. Each user's pictures are aligned (to extract the face) and split into a training set (pictures that are posted online and scraped) and a test set, at a **70%-30%** split.

For the reproduction, there are 523 users and 27,137 images, in the same train-test split.

The evaluation metric used is the error rate also called the protection rate.



FaceScrub

A Dataset With Over 100,000 Face Images of 530 People

Large face datasets are important for advancing face recognition research, but they are tedious to build, because a lot of work has to go into cleaning the huge amount of raw data. To facilitate this task, we developed an approach to building face datasets that detects faces in images returned from searches for public figures on the Internet, followed by automatically discarding those not belonging to each queried person.

The *FaceScrub* dataset was created using this approach, followed by manually checking and cleaning the results. It comprises a total of 106,863 face images* of male and female 530 celebrities, with about 200 images per person. As such, it is one of the largest public face databases. The dataset was also used as part of the MegaFace face recognition challenges.

The images were retrieved from the Internet and are taken under real-world situations (uncontrolled conditions). Name and gender annotations of the faces are included.

FaceScrub	Male	Female	Total
# people:	265	265	530
# images:	55,306	51,557	106,863

The dataset is released under a creative commons license. Note that we can only provide the URLs to the images (plus annotations), as we do not own the content (more details in the readme file). Please fill out this form to receive the access instructions.

Source: vintage.winklerbros.net. (n.d.). *vintage - resources*. [online] Available at: https://vintage.winklerbros.net/facescrub.html [Accessed 12 Dec. 2023].

Attack Setup - Perturbing User Images with the three attacks

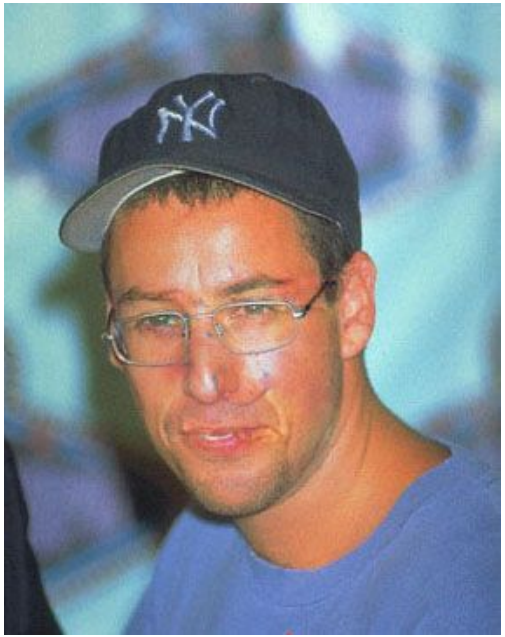
Fawkes 0.3

Fawkes 1.0

Lowkey

Perturbed Images with Fawkes 0.3 (High Mode)

+
Cloak
Perturbation



Perturbed Images with Fawkes 1.0 (High Mode)

+
Cloak
Perturbation

Perturbed Images with Lowkey

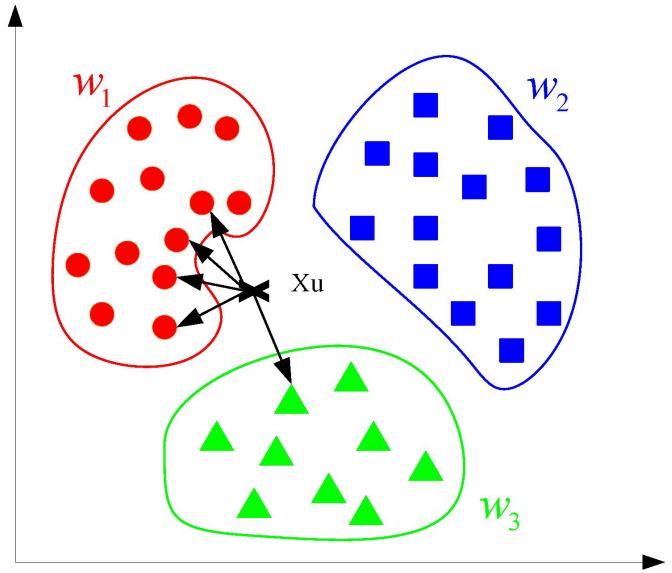
+
Cloak
Perturbation

Training Model

Standard Facial Recognition approach.
A model trainer trains a pre-trained feature extractor g(x).
Converts pictures into embeddings.
And evaluates using a 1-Nearest Neighbor rule.

$$x' = \underset{x'}{\operatorname{argmin}} \|g(x) - g(x')\|_2$$

And returns the identity *y'* associated with *x'*.



For a linear approach, a model trainer converts the feature extractor g(x) into a supervised classifier (by adding a linear layer on top of g) and fine-tunes the classifier on labeled pictures.

Feature Extractors

1. **FaceNet:** An Inception ResNet model pre-trained on VGG-Face2.
2. **WebFace:** An Inception ResNet model pre-trained on CASIA-WebFace. This feature extractor is used as a surrogate model in the Fawkes v0.3 attack.
3. **VGG-Face:** A VGG-16 model pre-trained on VGG-Face2.
4. **Celeb1M:** A ResNet trained on MS-Celeb-1M with the ArcFace loss. This feature extractor is used as a surrogate model in the Fawkes v1.0 attack.
5. **ArcFace:** A ResNet trained on CASIA-Webface with the ArcFace loss.
6. **MagFace:** A ResNet trained on MS-Celeb-1M with the MagFace loss.
7. **CLIP:** OpenAI's CLIP vision transformer model, which we fine-tuned on CASIA-WebFace and VGG-Face2.

Calculation of the Evaluation Metric

1. Test Accuracy:

$$Test\,acc = \frac{Number\,of\,correct\,predictions\,on\,the\,test\,set}{Total\,number\,of\,test\,samples} \cdot 100$$

2. Train Accuracy (User Cloaked):

$$Train\,acc(user\,cloaked) = \frac{Number\,of\,correct\,predictions\,on\,the\,cloaked\,training\,set}{Total\,number\,of\,cloaked\,training\,samples} \cdot 100$$

3. Test Accuracy (User Cloaked):

$$Test\,acc(user\,cloaked) = \frac{Number\,of\,correct\,predictions\,on\,the\,uncloaked\,test\,set}{Total\,number\,of\,uncloaked\,test\,samples} \cdot 100$$

4. Protection Rate (Error Rate):

$$Protectionrate = 1 - \frac{Number\,of\,correct\,predictions\,on\,the\,uncloaked\,test\,set}{Total\,number\,of\,uncloaked\,test\,samples} \cdot 100$$

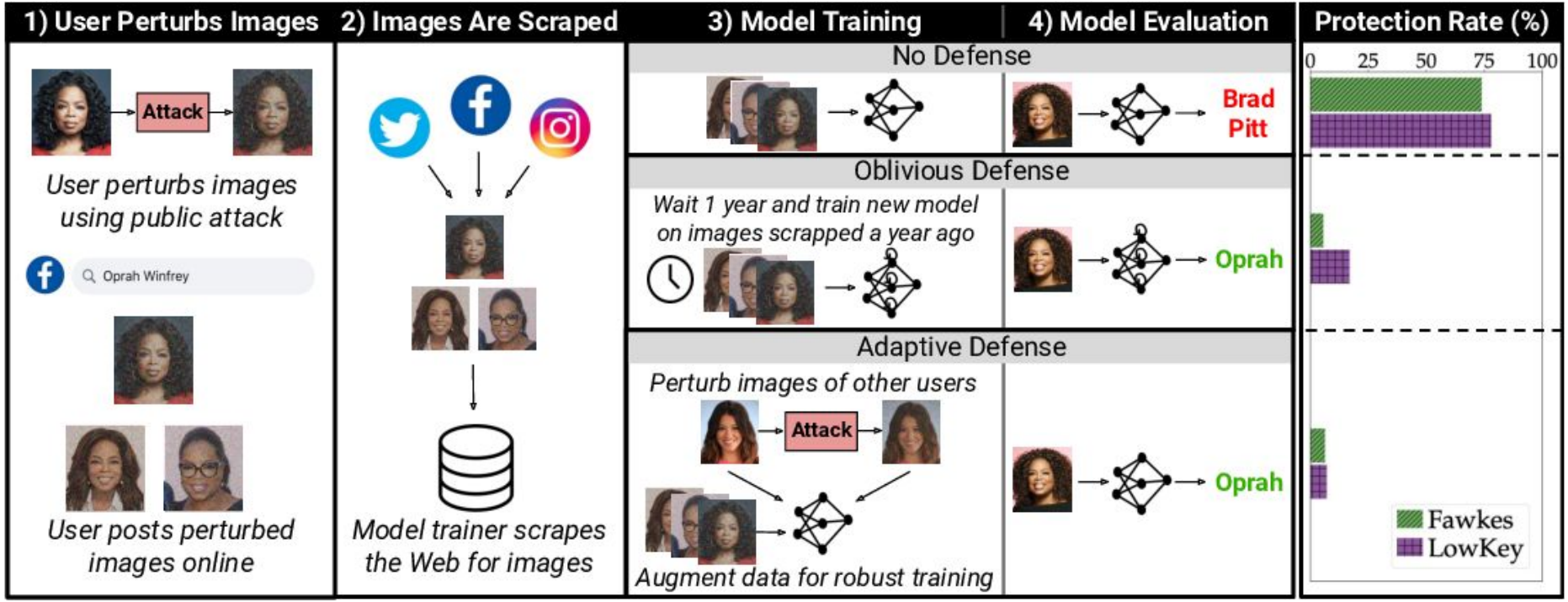
Defense Setup - Defense Strategies Evaluated against Fawkes and Lowkey

Baseline

Adaptive

Oblivious

Defense Strategies

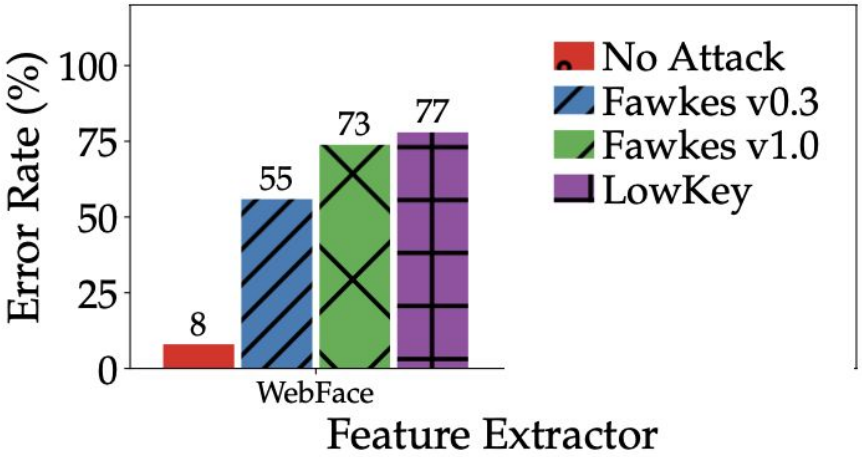


Source: Radiya-Dixit, E., Hong, S., Google, N. and Tramèr, F. (n.d.). *Published as a conference paper at ICLR 2022 DATA POISONING WON'T SAVE YOU FROM FACIAL RECOGNITION*. [online] Available at: https://openreview.net/pdf?id=B5XahNLmna#:~:text=With%20no%20defense%20strategy%2C%20the [Accessed 12 Dec. 2023].

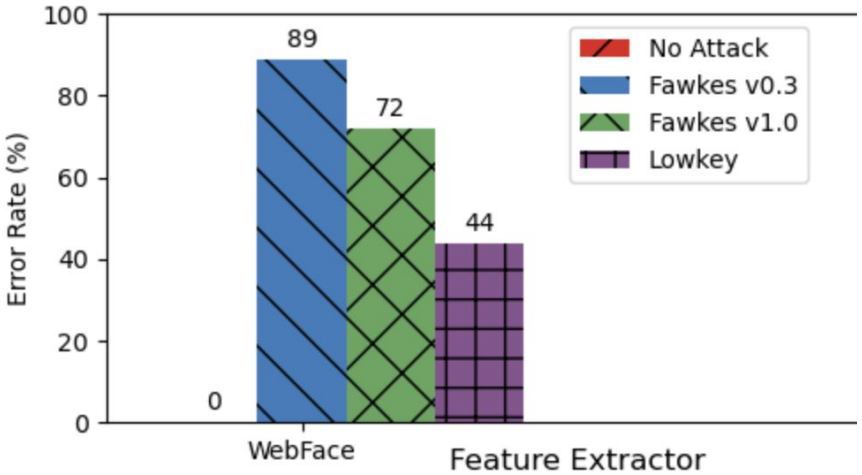
Baseline Defenses

- As a baseline defense, a model trainer collects perturbed pictures and trains a standard facial recognition model.
- In the paper, all attacks are evaluated against a non-robust **WebFace model.**
- The attacks are effective in this setting.
- For users who poisoned their online pictures, the model's error rate is 55-77%.
- For unprotected users, it is only only 8%.

Baseline Defenses



Original Results

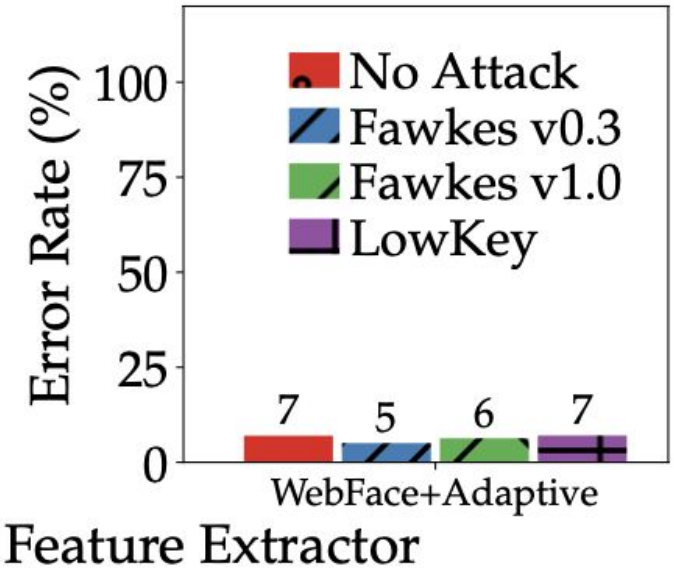


Reproduced Results

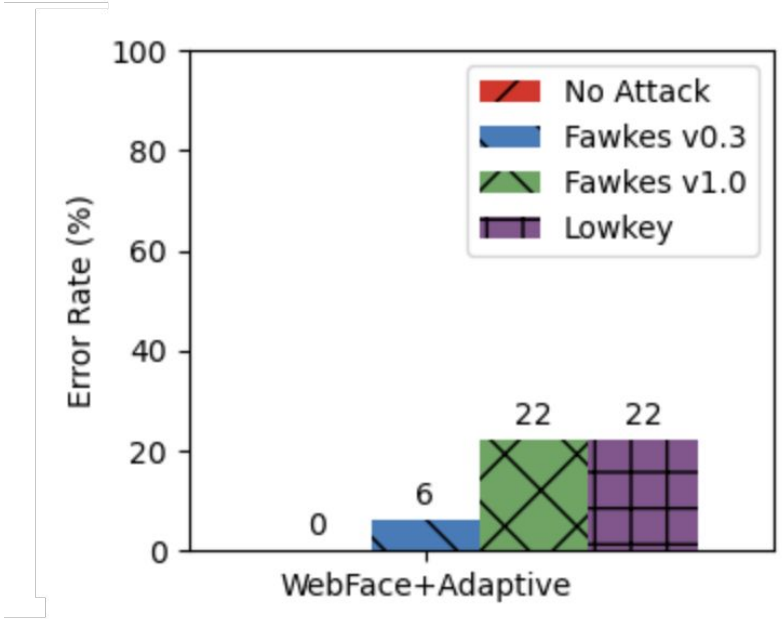
Adaptive Defenses

- As an adaptive defense, a model trainer uses the same attack as users (as a black-box) to build a training dataset augmented with perturbed pictures, and trains a model that is robust to the attack.
- Collects a public dataset (Xpublic, Ypublic) of unperturbed labeled faces.
  - half of the FaceScrub users in this case.
- Applies an attack (black-box) to perturb the images in the public dataset (Xpublic(adv) ←Attack(Xpublic)).
  - Fawkes and LowKey attacks to obtain perturbed samples.
- Trains a model using both unperturbed and perturbed samples and aim to generate similar embeddings for unperturbed and perturbed pictures of the same user.
  - Robustly fine-tune the pre-trained WebFace feature extractor.
- Teaches the model to produce robust embeddings that resist perturbations and generalizes the learned robust features to handle perturbations on other users' pictures.
- All attempts to attack in this case are unsuccessful.
- The robust model has the same error rate for users trying to attack as it does for the average user.

Adaptive Defenses



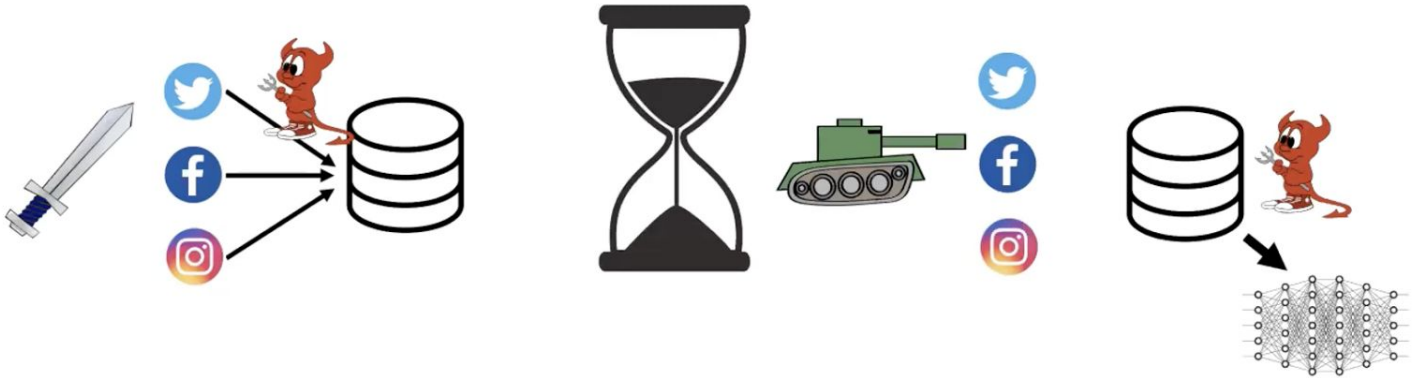
Original Results



Reproduced Results

Oblivious Defenses

New models can be applied *retroactively*.



Model trainer scrapes pictures produced with weak attack

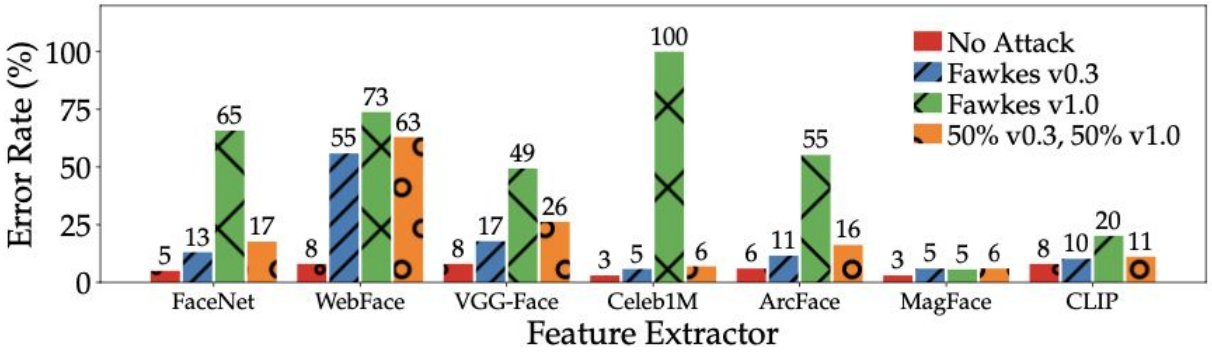
Users switch to stronger attacks, but new model can be trained solely on data collected in the past

Oblivious Defenses

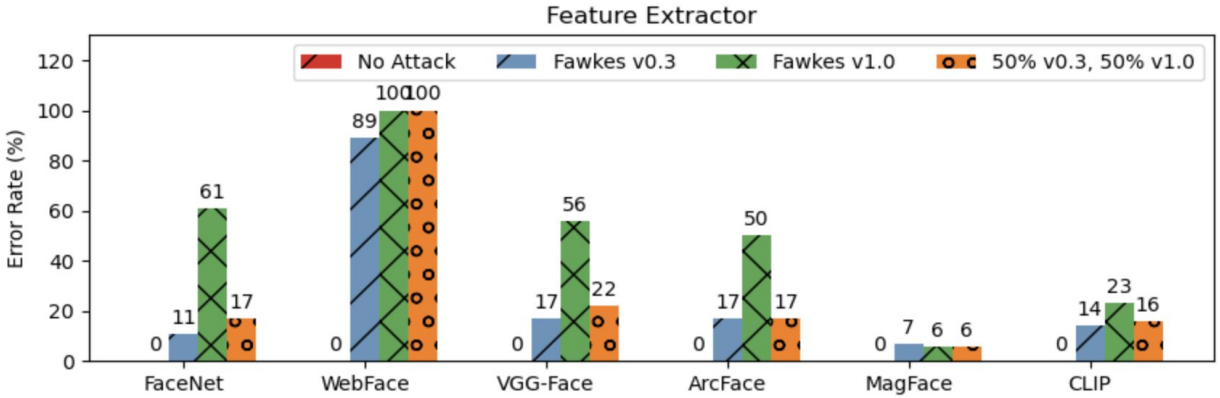
- As an oblivious defense, a model trainer collects perturbed pictures, waits until a new facial recognition model is released, and uses the new model to nullify the protection of previously collected pictures.
- This oblivious strategy demonstrates the futility of preventing facial recognition with data poisoning, so long as progress in facial recognition models is expected to continue in the future.
- To bypass this oblivious defense strategy, a poisoning attack must fool not only today's models, but also all future models.

Oblivious Defenses - Fawkes Attacks

Original Results

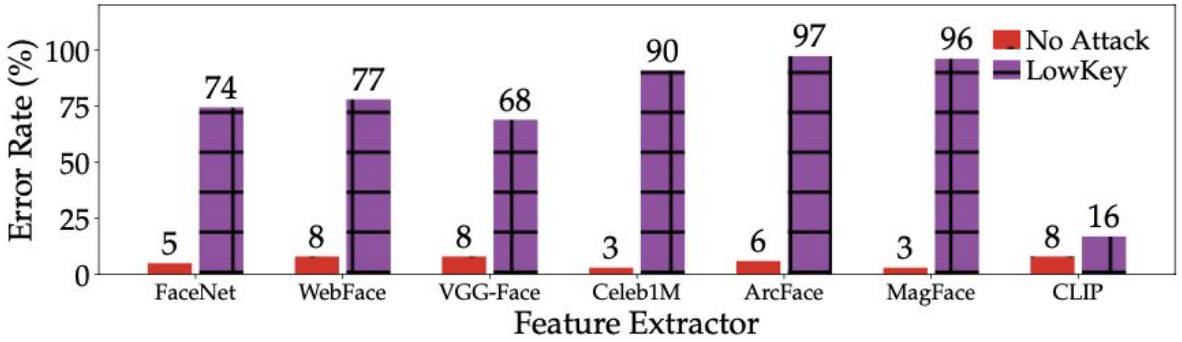


Reproduced Results

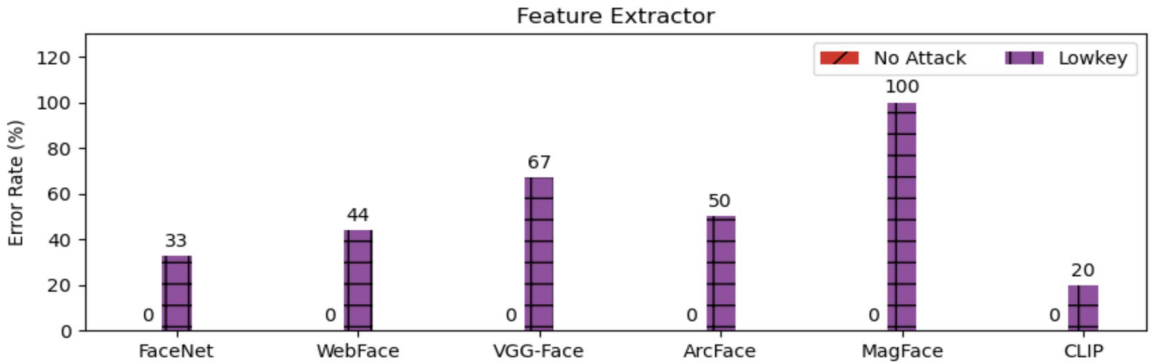


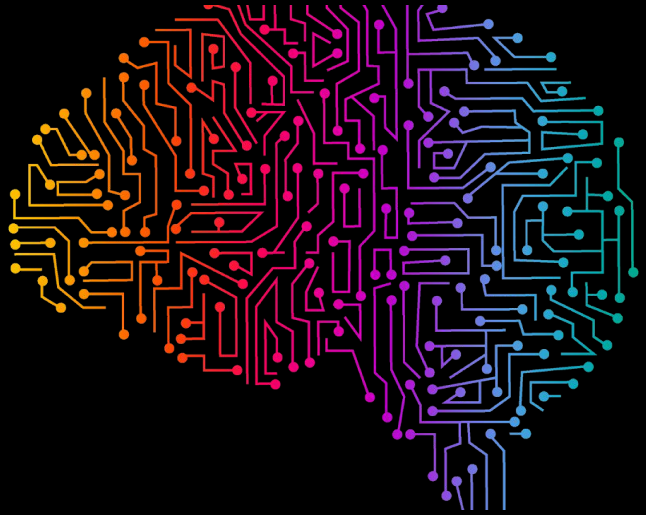
Oblivious Defenses - Lowkey Attack

Original Results



Reproduced Results

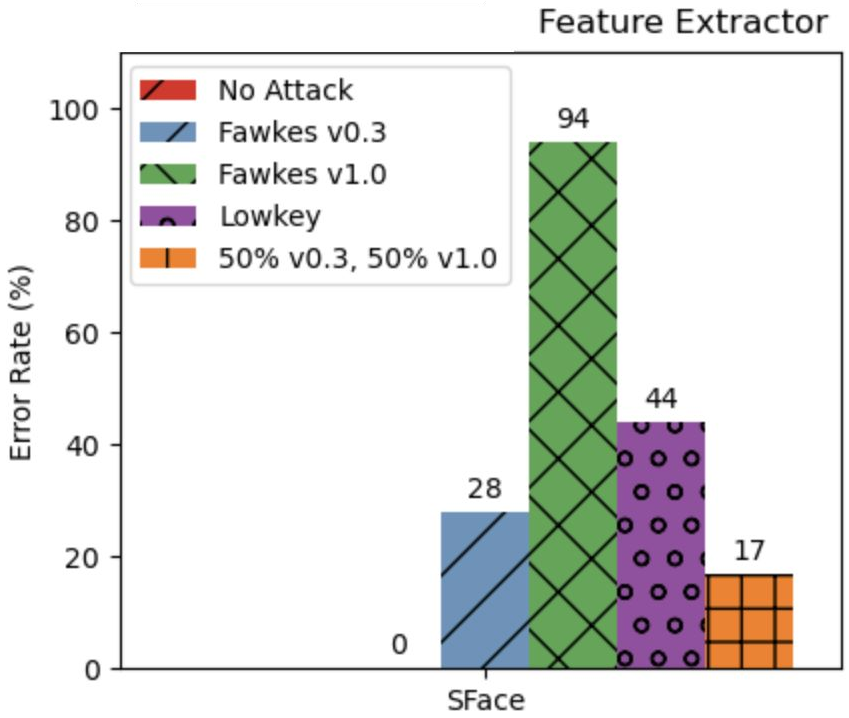




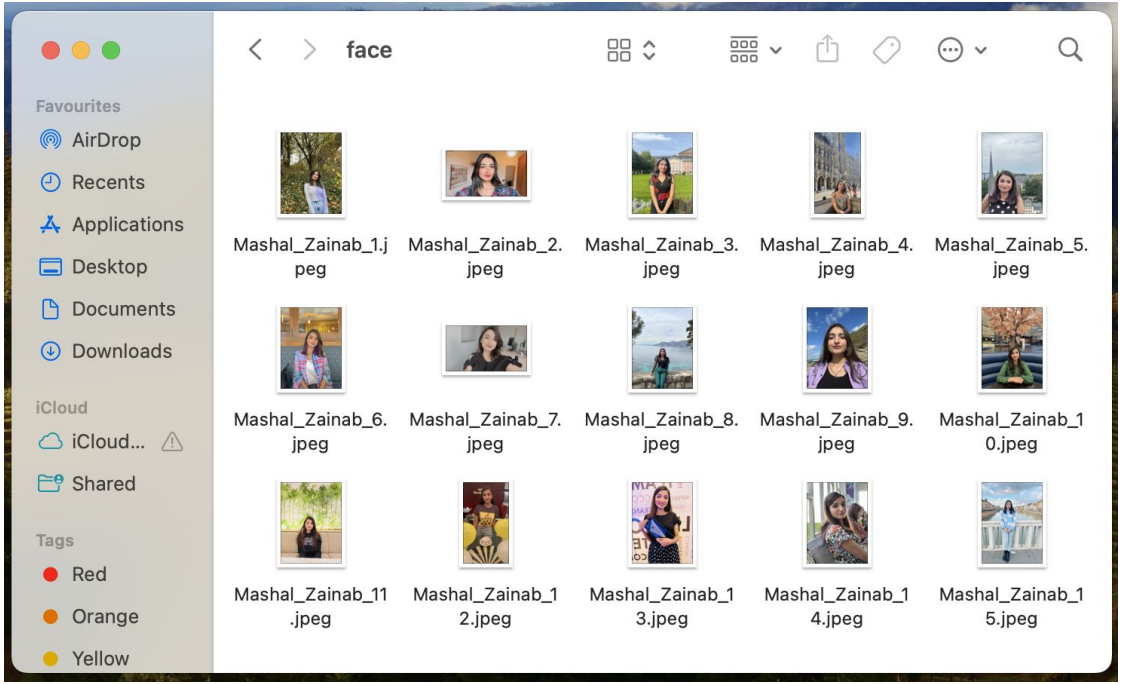
Part 3: Replication of the Experiments

Additional Trained Model SFace

- SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data.
- It is generated using a class-conditional generative adversarial network and is designed to address privacy and ethical concerns surrounding public face recognition datasets.
- Trained using the Deepface library.
- For other models, Lowkey was performing better but for this one it does not.

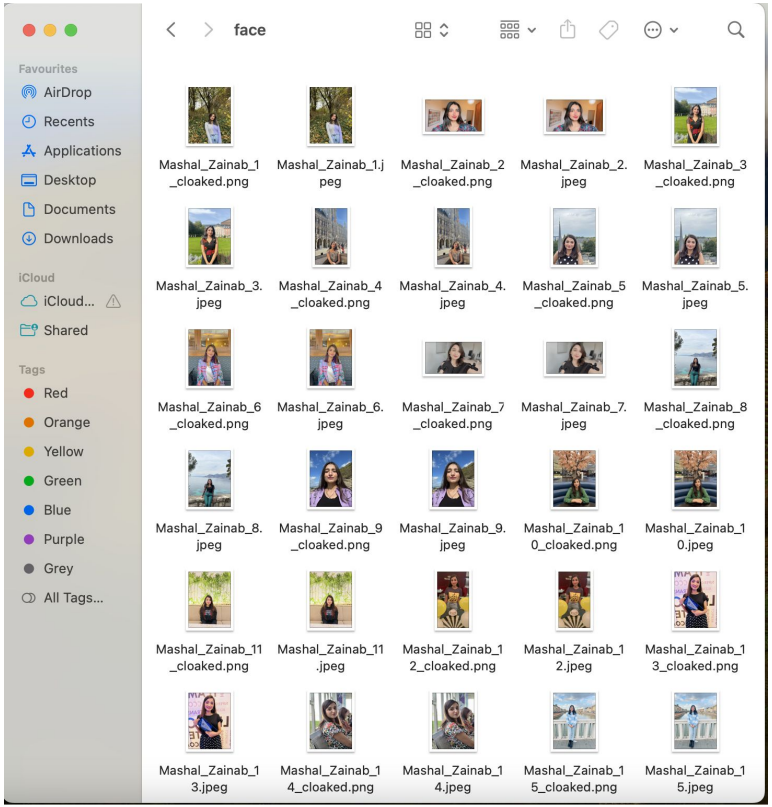


Becoming the Attacking user

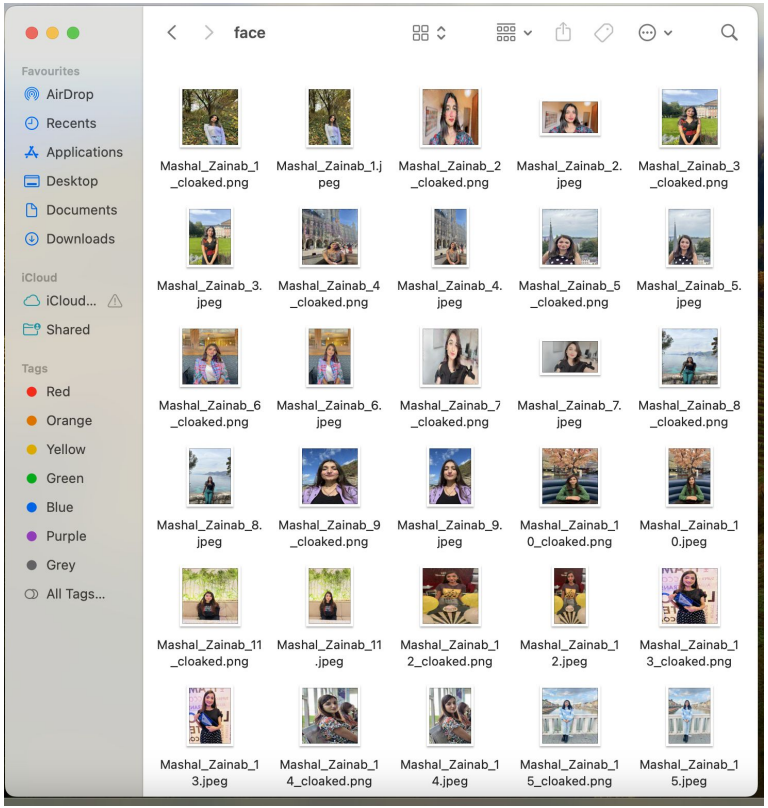


Perturbed 15 pictures from different years.

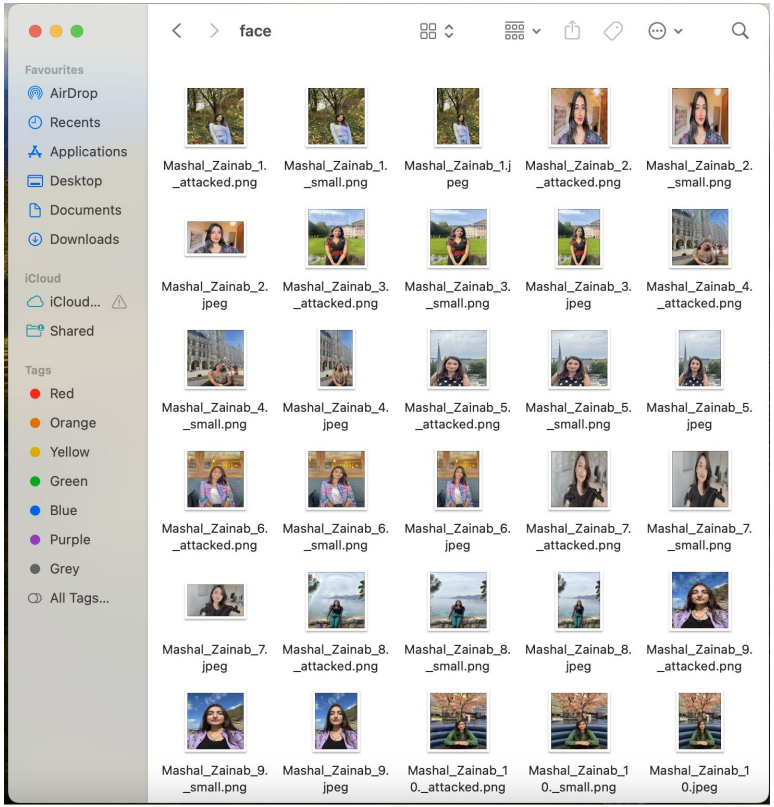
Perturbed Images with Fawkes 0.3



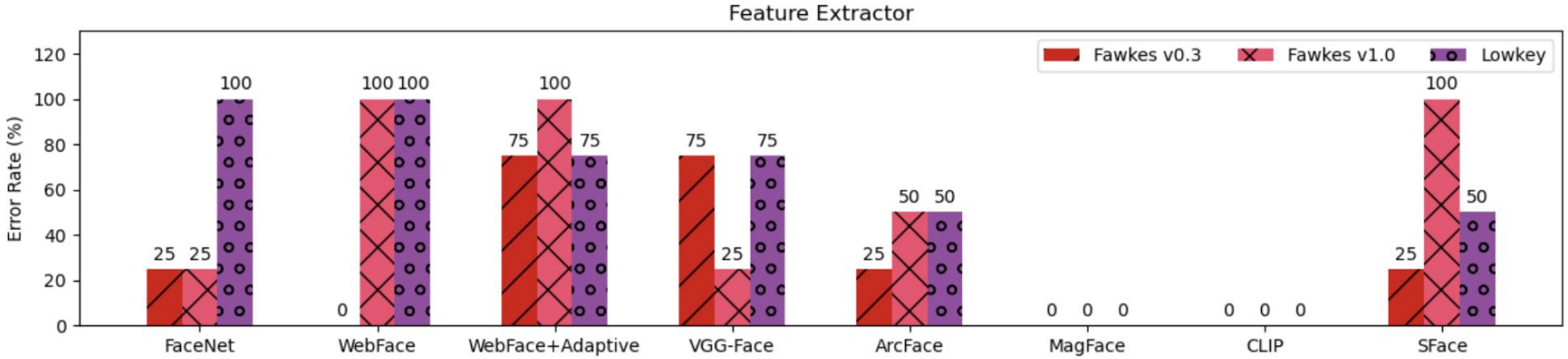
Perturbed Images with Fawkes 1.0



Perturbed Images with Lowkey



Defenses


Feature Extractor
Error Rate (%)
Fawkes v0.3
Fawkes v1.0
Lowkey
FaceNet
WebFace
WebFace+Adaptive
VGG-Face
ArcFace
MagFace
CLIP
SFace

Conclusion

- Data poisoning is not an effective defense against facial recognition models trained on web-scraped pictures.
- Future developments in computer vision can easily nullify the protection of pictures collected in the past.
- There is no "arms race" between attackers and defenders in facial recognition poisoning.
- Users cannot rely on data poisoning to protect their privacy in the context of facial recognition technology.

References

- Boutros, F. (2023). *fdbtrs/SFace-Privacy-friendly-and-Accurate-Face-Recognition-using-Synthetic-Data*. [online] GitHub. Available at: https://github.com/fdbtrs/SFace-Privacy-friendly-and-Accurate-Face-Recognition-using-Synthetic-Data [Accessed 17 Dec. 2023].
- Boutros, F., Huber, M., Siebke, P., Rieber, T. and Damer, N. (n.d.). *SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data*. [online] Available at: https://arxiv.org/pdf/2206.10520.pdf [Accessed 17 Dec. 2023].
- Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J.P., Taylor, G. and Goldstein, T. (2020). *LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition*. [online] openreview.net. Available at: https://openreview.net/forum?id=hJmtwocEqzc [Accessed 17 Dec. 2023].
- GitHub. (n.d.). *Release Fawkes 0.3 · Shawn-Shan/fawkes*. [online] Available at: https://github.com/Shawn-Shan/fawkes/releases/tag/v0.3 [Accessed 17 Dec. 2023].
- lowkey.umiacs.umd.edu. (n.d.). *Lowkey*. [online] Available at: https://lowkey.umiacs.umd.edu/.
- Meng, Q. (2023). *MagFace*. [online] GitHub. Available at: https://github.com/IrvingMeng/MagFace/ [Accessed 17 Dec. 2023].
- OpenAI (2022). *CLIP*. [online] GitHub. Available at: https://github.com/openai/CLIP .
- Serengil, S.I. (2023). *deepface*. [online] GitHub. Available at: https://github.com/serengil/deepface?tab=readme-ov-file [Accessed 17 Dec. 2023].
- Shan, S. (2023). *Shawn-Shan/fawkes*. [online] GitHub. Available at: https://github.com/Shawn-Shan/fawkes [Accessed 17 Dec. 2023].
- Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H. and Zhao, B. (2020). *Open access to the Proceedings of the 29th USENIX Security Symposium is sponsored by USENIX. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models Fawkes: Protecting Privacy against Unauthorized Deep Learning Models*. [online] Available at: https://www.shawnshan.com/files/publication/fawkes.pdf [Accessed 17 Dec. 2023].
- Tramer, F. (2023). *FaceCure*. [online] GitHub. Available at: https://github.com/ftramer/FaceCure/tree/main?tab=readme-ov-file [Accessed 17 Dec. 2023].
- Uchicago.edu. (2020). *Fawkes*. [online] Available at: https://sandlab.cs.uchicago.edu/fawkes/ .
- vintage.winklerbros.net. (n.d.). *vintage - resources*. [online] Available at: https://vintage.winklerbros.net/facescrub.html.

Thank you!

Questions are appreciated!

https://github.com/mashalbhatti/AI_and_Cybersecurity_Project_FaceCure