Fawkes: Protecting Privacy against Unauthorized Deep Learning Models

Shawn Shan†
Emily Wenger†
Jiayun Zhang
Huiying Li
Haitao Zheng
Ben Y. Zhao

† denotes co-first authors with equal contribution
Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models
Facial Recognition Models are Easy to Build

Less time to train larger/more powerful models

Cheaper, faster hardware
Facial Recognition Models are Easy to Build

- Less time to train larger/more powerful models
- Cheaper, faster hardware
- Labelled training data everywhere
Facial Recognition Models are Easy to Build

- Less time to train larger/more powerful models
- Cheaper, faster hardware
- Labelled training data everywhere

Anyone with limited coding knowledge and computational power can train powerful facial recognition models
Facial Recognition Models are Easy to Build

- Less time to train larger/more powerful models
- Cheaper, faster hardware
- Labelled training data everywhere

- Anyone with limited coding knowledge and computational power can train powerful facial recognition models

But what if the **wrong** people take advantage of this new accessibility?
Personal Images Co-opted to Train Facial Recognition Models
Personal Images Co-opted to Train Facial Recognition Models
Facial Recognition Models Easily Misused
Facial Recognition Models Easily Misused

This is Emily W.
Here are more pics and Google results

Facial ID Service
Facial Recognition Models Easily Misused

Malicious entity -> Facial ID Service

This is Emily W. Here are more pics and Google results

Location info: stalking
Facial Recognition Models Easily Misused

Malicious entity

Facial ID Service

This is Emily W. Here are more pics and Google results

Location info: stalking

Family info: phishing/extortion
Facial Recognition Models Easily Misused

Malicious entity

Facial ID Service

This is Emily W. Here are more pics and Google results

Location info: stalking

Family info: phishing/extortion

Personal info: employment decisions
Facial Recognition Models Easily Misused

Malicious entity

Facial ID Service

This is Emily W. Here are more pics and Google results

Other info could lead to:
- Racial discrimination
- Political oppression
- Religious persecution

- Location info: stalking
- Family info: phishing/extortion
- Personal info: employment decisions
That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

...
That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.

Database of 3B scraped images
That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.

Diagram:
- A Single User Image
- Clearview.AI
- Database of 3B scraped images
- User’s other images online
- Personal Information
That Reality is Here, Today

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.

Known Clearview.ai customers include government agencies, law enforcement departments, and private citizens.

A Single User Image

User’s other images online

Personal Information

Database of 3B scraped images
In This Talk

**Fawkes:**
*Privacy armor* that protects privacy by preventing your images from being used to train ML models against you.

- Fawkes Design
- Evaluation
- Live Tests against Face Recognition Services
Goals and Assumptions

User  Tracker (e.g. Clearview)
Goals and Assumptions

User

Tracker (e.g. Clearview)

Limited computational resources
Goals and Assumptions

User

Tracker (e.g. Clearview)

Limited computational resources + Well-trained feature extractor
Goals and Assumptions

User

Limited computational resources

Well-trained feature extractor

Goal: Protected images posted online

Tracker (e.g. Clearview)
Goals and Assumptions

User Tracker (e.g. Clearview)

Limited computational resources
Well-trained feature extractor
Goal: Protected images posted online
Tracker: Protected images scraped from online
Goals and Assumptions

User Tracker (e.g., Clearview)

Limited computational resources + Well-trained feature extractor

Goal: Protected images posted online

Protected images scraped from online + Extensive computational resources

Limited computational resources

Well-trained feature extractor

Goal: Protected images posted online

Protected images scraped from online

Extensive computational resources
Goals and Assumptions

User

Limited computational resources
Well-trained feature extractor

Goal: Protected images posted online

Tracker (e.g. Clearview)

Protected images scraped from online
Extensive computational resources

Goal: Effective facial recognition model
Intuitive View of Facial Recognition Models

Training Images

[Images of individuals]
Intuitive View of Facial Recognition Models

Training Images → Feature extractor
Intuitive View of Facial Recognition Models

Training Images

Feature extractor

Feature Representation

$x_2$

$x_1$
Intuitive View of Facial Recognition Models

Training Images

Feature extractor

Feature Representation

...
Intuitive View of Facial Recognition Models

Training Images

Feature extractor

Test input

Decision boundary

Feature Representation
Intuitive View of Facial Recognition Models

Training Images

Feature extractor

Feature Representation

Test input

Decision boundary
Intuitive View of Facial Recognition Models

Training Images

Feature extractor

Feature Representation

Classification based on feature space separation

Test input

Ben Zhao

Heather Zheng

Beyoncé

Emily

Decision boundary
To evade unwanted facial recognition, change the feature space representation of user images.
Key Intuition for Fawkes

To evade unwanted facial recognition, change the feature space representation of user images.

**Intuition:** Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.
Key Intuition for Fawkes

To evade unwanted facial recognition, change the feature space representation of user images.

**Intuition:** Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.
Key Intuition for Fawkes

To evade unwanted facial recognition, change the feature space representation of user images.

**Intuition:** Emily's cloaked images are perturbed to have similar feature space representations to Beyoncé's images, distinct from Emily's original feature space representation.
How to Generate Cloak?

Compute cloak perturbation ($\Delta$) by solving an optimization problem

- Goal: mimic feature representations of target class $T$
- Constraint: perturbation should be indistinguishable by humans
How to Generate Cloak?

Compute cloak perturbation ($\Delta$) by solving an optimization problem

- **Goal:** mimic feature representations of target class $T$
- **Constraint:** perturbation should be indistinguishable by humans

$\Phi(X)$: internal representation in feature extractor $\Phi$

$$\min \ \text{Distance}(\Phi(X_s + \Delta), \Phi(X_t))$$

Minimize $L2$ distance between internal representations
How to Generate Cloak?

Compute cloak perturbation (Δ) by solving an optimization problem

- **Goal**: mimic feature representations of target class T
- **Constraint**: perturbation should be indistinguishable by humans

\[
\Phi(X): \text{ internal representation in feature extractor } \Phi
\]

\[
\min \ Distance(\Phi(X_s + \Delta), \Phi(X_t))
\]

\[
s.t. \quad \text{perturb\_magnitude}(X_s + \Delta, X_s) < P_{budget}
\]
How to Generate Cloak?

Compute cloak perturbation ($\Delta$) by solving an optimization problem

- Goal: mimic feature representations of target class $T$
- Constraint: perturbation should be indistinguishable by humans

$$\Phi(X): \text{internal representation in feature extractor } \Phi$$

$$\min \ Distance(\Phi(X_s + \Delta), \Phi(X_t))$$

$$s.t. \ \text{perturb\_magnitude}(X_s + \Delta, X_s) < P_{budget}$$

$DSSIM$: an objective measure for image distortion

Minimize $L2$ distance between internal representations
Protection under Baseline Conditions

Original Images + Well-trained feature extractor = Tracker
Protection under Baseline Conditions

User + Fawkes

Tracker

Original Images

Well-trained feature extractor
Protection under Baseline Conditions

User + Fawkes

Tracker

Original Images

Well-trained feature extractor

Cloaked Images
Protection under Baseline Conditions

User + Fawkes

Tracker

Original Images

Well-trained feature extractor

Cloaked Images

Well-trained feature extractor

Face recognition model
Protection under Baseline Conditions

**User + Fawkes**

Original Images

Well-trained feature extractor

**Tracker**

Cloaked Images

Well-trained feature extractor

Face recognition model

**Protection Success Rate:** Percentage of real (unmodified) user images misclassified by tracker’s model
Protection under Baseline Conditions

User + Fawkes

Tracker

<table>
<thead>
<tr>
<th>Feature Extractor Used</th>
<th>Protection Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGFace2 + InceptionResNet</td>
<td>100%</td>
</tr>
<tr>
<td>VGGFace2 + DenseNet</td>
<td>100%</td>
</tr>
<tr>
<td>WebFace + InceptionResNet</td>
<td>100%</td>
</tr>
<tr>
<td>WebFace + DenseNet</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Protection Success Rate**: Percentage of real (unmodified) user images misclassified by tracker’s model.
Fawkes knows tracker’s FE, uses it to compute cloak

Protection Rate: 100%
Protection under Realistic Conditions

**Known Feature Extractor**
Fawkes knows tracker's FE, uses it to compute cloak

**Protection Rate:** 100%

**Transferability:** models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.
Protection under Realistic Conditions

**Known Feature Extractor**
Fawkes knows tracker’s FE, uses it to compute cloak

Protection Rate: 100%

**Unknown Feature Extractor**
Tracker uses unknown FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

**Transferability:** models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.
Protection under Realistic Conditions

**Known Feature Extractor**
Fawkes knows tracker's FE, uses it to compute cloak

Protection Rate: 100%

**Unknown Feature Extractor**
Tracker uses unknown FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

**Train from scratch**
Tracker does not use FE. Fawkes computes cloak on local FE & relies on transferability

Protection Rate: >95%

Transferability: models trained on different data (but same application domain) often share similarity in feature space representation, so effects of perturbations from one can transfer to a different feature extractor or model.
Protection against State of the Art APIs
Protection against State of the Art APIs

• How well does Fawkes work on real world Face recognition APIs?
Protection against State of the Art APIs

- How well does Fawkes work on real world Face recognition APIs?
Protection against State of the Art APIs

• How well does Fawkes work on real world Face recognition APIs?

1. Train facial recognition model on public API
2. Training data includes 1 cloaked user X (all their images are cloaked by Fawkes Using existing feature extractor)
3. Test result model with uncloaked images of user X
Protection against State of the Art APIs

• How well does Fawkes work on real world Face recognition APIs?

<table>
<thead>
<tr>
<th>Face Recognition API</th>
<th>Protection Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Protection</td>
</tr>
<tr>
<td>AWS Rekognition</td>
<td>0%</td>
</tr>
<tr>
<td>Microsoft Azure</td>
<td>0%</td>
</tr>
<tr>
<td>Face++</td>
<td>0%</td>
</tr>
</tbody>
</table>

• Result is 100% success (no clean images identified as the user, all misclassified)

1. Train facial recognition model on public API
2. Training data includes 1 cloaked user X (all their images are cloaked by Fawkes using existing feature extractor)
3. Test result model with uncloaked images of user X
Even More Challenging Real-World Scenarios
Even More Challenging Real-World Scenarios

• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
Even More Challenging Real-World Scenarios

• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  • Cloaking can succeed if cloaked images outnumber uncloaked images
  • Sybil accounts help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)
• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  • Cloaking can succeed if cloaked images outnumber uncloaked images
  • Sybil accounts help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)

• What if the tracker tries to detect/remove cloaked effects?
  • Using tools like image transformation, anomaly detection
  • Ineffective against Fawkes
Even More Challenging Real-World Scenarios

• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  • Cloaking can succeed if cloaked images outnumber uncloaked images
  • *Sybil accounts* help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)

• What if the tracker tries to detect/remove cloaked effects?
  • Using tools like image transformation, anomaly detection
  • Ineffective against Fawkes

• Limitations of Fawkes
• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  • Cloaking can succeed if cloaked images outnumber uncloaked images
  • Sybil accounts help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)

• What if the tracker tries to detect/remove cloaked effects?
  • Using tools like image transformation, anomaly detection
  • Ineffective against Fawkes

• Limitations of Fawkes
  • Not guaranteed to be effective against future models
  • Only a tip of the iceberg
• What if the tracker has original, uncloaked images?
  • Pre-Fawkes images, public sources (newspapers, company page), images shared by friends
  • Cloaking can succeed if cloaked images outnumber uncloaked images
  • Sybil accounts help: boost protection from 30% to 95% (1:1 cloak:uncloaked ratio)

• What if the tracker tries to detect/remove cloaked effects?
  • Using tools like image transformation, anomaly detection
  • Ineffective against Fawkes

• Limitations of Fawkes
  • Not guaranteed to be effective against future models
  • Only a tip of the iceberg

• More details in paper!
Thank You!
Thank You!

- More on http://sandlab.cs.uchicago.edu/fawkes
- Source code
- Binaries for MacOS/Windows/Linux
- FAQs

- Encouraging initial response from users
- 2.5K downloads as of July 20th