On the Robustness of Facial Privacy Protection Fawkes against AI Denoising Attacks

Tong Zhou
Department of Computer Science
Virginia Tech

Xinbei Zhu
Department of Computer Science
Virginia Tech

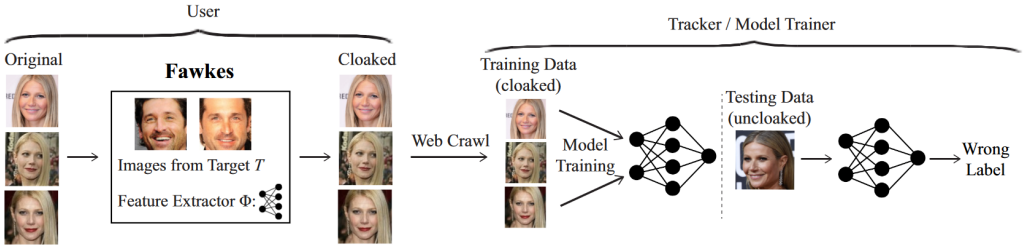
Abstract

This study addresses the growing privacy concerns stemming from the unauthorized use of publicly available photos for facial recognition training by third parties. Focusing on the Fawkes algorithm, developed by Shan et al. at the University of Chicago's Sand Lab, we examine an approach to counteract its effects.
Our project aims to test the robustness of these perturbations against advanced denoising techniques. The core of our study involves a comparative analysis of three distinct facial recognition models: one trained on standard internet-crawled images, another on Fawkes-cloaked images, and a third on images post-CNN denoising. This practical approach ensures that our findings are not confined to theoretical but are tested in real-world scenarios. Our findings demonstrate the feasibility of this approach, revealing that CNN models are effective at countering the privacy-preserving noise introduced by Fawkes.

1 Introduction

1.1 Background

Figure 1: The functionality of the Fawkes



The rapid advancement of machine learning technology brings significant privacy concerns, particularly in facial recognition. A New York Times article by Kashmir Hill highlighted Clearview.ai, a company that compiled over 3 billion photos from the Internet and social media to create facial recognition models without public consent. This underscores the ease of creating intrusive monitoring tools via deep learning. In response, the University of Chicago's SAND lab developed Fawkes, a system allowing individuals to protect their images against such unauthorized use. Fawkes subtly alters photos ("cloaks") without noticeable distortion, effectively countering advanced facial recognition technologies. The functionality of the Fawkes model is illustrated in Figure 1.

1.2 Motivation and Problem Description

While the Fawkes algorithm has been demonstrated in research papers to effectively safeguard image privacy, our study explores the possibility of circumventing this protection. Understanding the potential vulnerabilities of Fawkes is essential, especially in assessing its resilience against image restoration technologies. In our project, we specifically view the perturbations added by the Fawkes algorithm as a form of noise introduced into the images. Our approach involves simulating real-world scenarios to ensure the practical relevance of our findings. Our objective is to investigate whether these noise-like alterations can be effectively neutralized using advanced denoising models. By employing denoising techniques, we aim to strip away the cloaking effects that Fawkes adds, essentially 'cleaning' the images back to their original state.

2 Existing Approaches to Breaking Fawkes

The Fawkes paper discusses various potential countermeasures, yet none significantly impact the effectiveness of the Fawkes model. However, there is no existing method that specifically employs a denoising model to remove the "noises" introduced into images by Fawkes.

1. **Image Transformation**. This method modifies training images by augmentation, blurring, or adding noise, aiming to lessen the impact of minor image perturbations. Online image sharing, which typically includes compression, could also influence cloak effectiveness. Nonetheless, these techniques don't effectively overcome cloaks. The addition of Gaussian noise to images, for example, may affect overall classification accuracy but does not significantly reduce the high success rate of cloak protection, even as the noise's standard deviation increases.

2. **Robust Model**. The study evaluates cloaking effectiveness when the tracking model outperforms the user's feature extractor. In a test scenario, the cloak's privacy protection success rate fell to 64%. For those prioritizing privacy, making the cloak's alterations more noticeable enhances this rate. A cloak's distortion level above 0.01 DSSIM achieves a 100

3 Proposed Approach

Our experiment involves testing facial recognition models using three kinds of images: those found on the internet, images changed by Fawkes, and images that have been cleaned of noise. This test is key to see how strong Fawkes' hiding method is against powerful cleaning techniques. It also helps us understand the ongoing challenge of protecting personal privacy against the growing power of facial recognition technology.

3.1 Data Pre-processing

Based on the findings in the Fawkes paper, which indicate that increased label density enhances cloaking effectiveness, we have chosen to use 360 labels (categories) with approximately 80 images in each to improve the effectiveness of cloaking. We excluded categories with fewer than 20 images, amounting to 1.1% of the data. The dataset was then equally divided into two parts: one for the denoising model and the other for the facial recognition model. The facial recognition dataset was further split, half for training and half for testing.

To mirror real-world conditions in our experiment, we considered that not everyone might adopt Fawkes immediately. Therefore, it's realistic to assume that images sourced by third parties from the internet would be a mixture of original and Fawkes-cloaked images. To simulate this scenario, we split the testing dataset into two sections. One section was cloaked using Fawkes to introduce cloaks, while the other remained original. These two parts were then merged to create a mixed dataset of cloaked and uncloaked images, as depicted in our Figure 2. This approach aims to closely replicate the likely diversity of images in actual internet sources, ensuring our experiment aligns closely with potential real-world applications.

Figure 2: Flow Path of Data Pre-processing

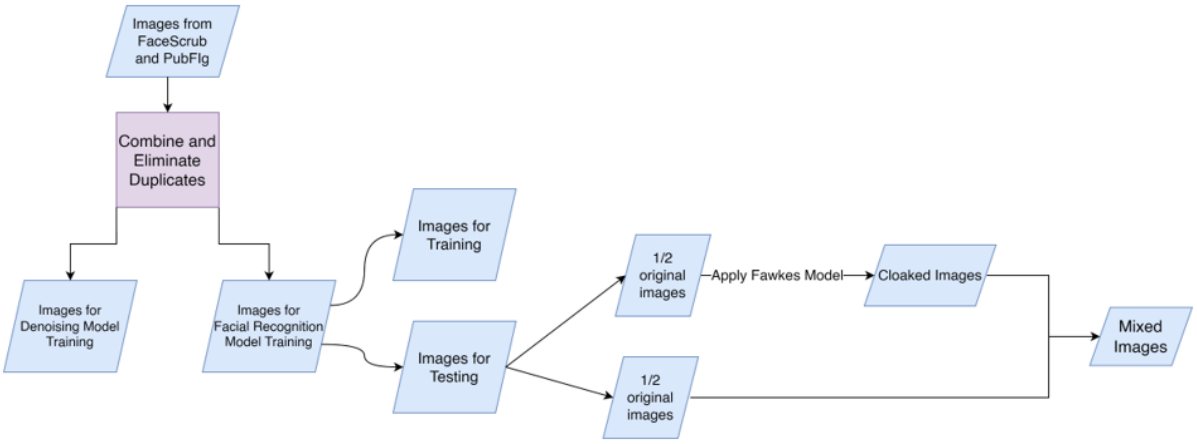
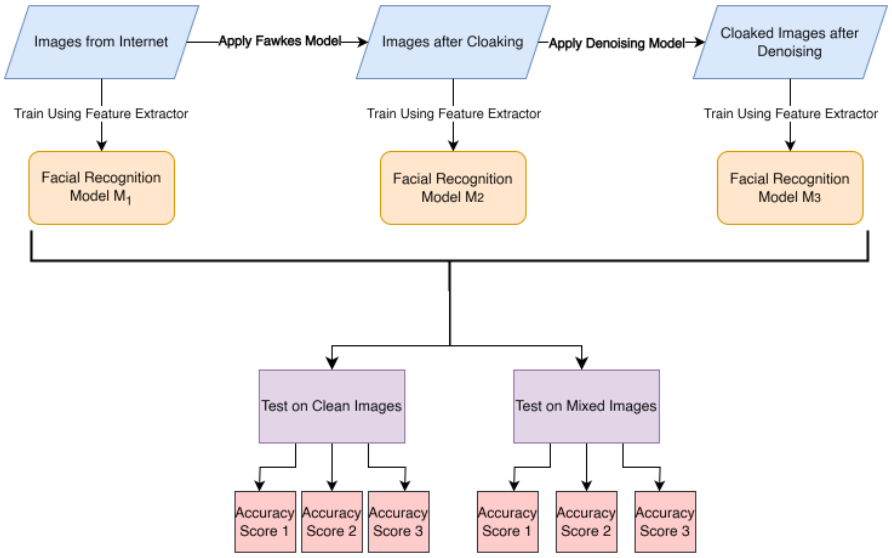


Figure 3: Flow Diagram of Methodology



3.2 Methodology

Our method follows a straightforward process, as shown in Figure 3. Initially, we collect images from the internet. These images are then processed through the Fawkes model, resulting in a set of cloaked images. Subsequently, we apply a denoising model to these cloaked images to produce a set of denoised images. Consequently, we develop three distinct facial recognition models, each trained on one of these three datasets using the same feature extractor: original internet-crawled images, cloaked images, and denoised images.

For evaluating the performance of these models, we use the accuracy rate as our primary metric. We test each of the three models on two types of datasets: clean (uncloaked) images and a mix of cloaked and uncloaked images. The rationale for testing on a mixed image dataset is to mimic real-world scenarios, where both cloaked and uncloaked images are likely to be present. Testing on a clean image dataset provides a more comprehensive analysis by offering a baseline comparison. This dual-testing approach allows us to thoroughly assess the performance of our facial recognition models in various situations, ensuring a well-rounded evaluation of their effectiveness.

Denoising Model. The initial step in our methodology involves training the denoising model. To do this, we first gather images from our selected datasets. These images are then processed through

Figure 4: Flow Diagram of Training Denoising Model

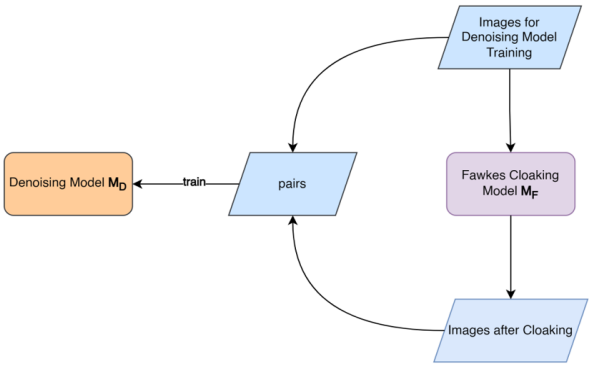
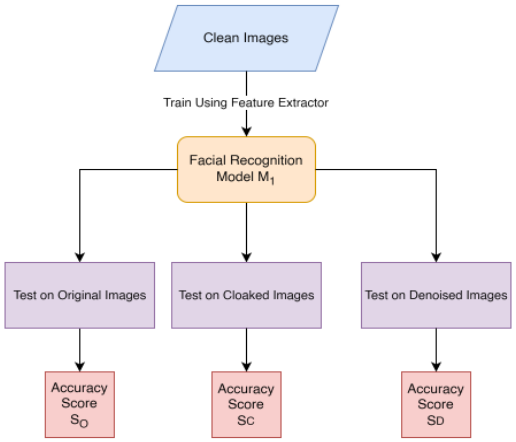


Figure 5: Midterm Flow Diagram



the Fawkes model, resulting in a set of cloaked, or as we refer to them, "noisy" images. This process provides us with pairs of images: the original ones and their cloaked counterparts. We use these paired images - original and noisy - as the training data for our denoising model. The flow diagram of training denoising model is shown in Figure 4. we have chosen to utilize a simple Convolutional Neural Network (CNN) as our denoising model. This decision is based on the CNN's proven capability in image processing tasks, including its effectiveness in identifying and filtering out noise from images.

Feature Extractor. For training our facial recognition models, we have selected two distinct feature extractors: ResNet and AlexNet. These choices are motivated by the unique strengths and architectures of each model. ResNet is expected to provide high accuracy and efficiency in facial recognition tasks. AlexNet, on the other hand, with its simpler architecture, offers a more straightforward approach and has been historically significant in the field of deep learning for image recognition. By employing both of these feature extractors, we can gain a broader and more in-depth perspective on the performance of our CNN denoising model.

4 Experimental Evaluation

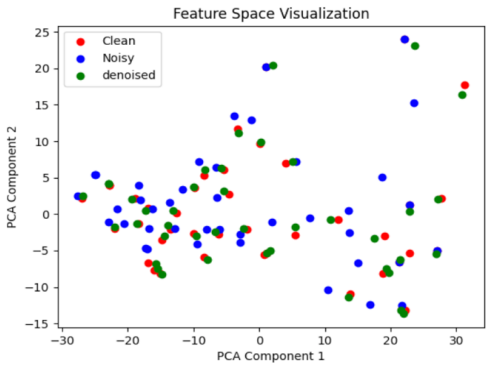
4.1 Initial Results (Midterm Results) Analysis

We conducted training on a ResNet model using clean images and then evaluated its performance with three different types of images: clean, cloaked, and denoised (Figure 5). The results were quite revealing. The model achieved an accuracy of 80.77% with clean images, serving as a baseline for comparison. When tested on cloaked images, the accuracy fell to 70.65%, underscoring the difficulty

Table 1: Initial (Midterm) Result

Test Set	Accuracy
Clean Images	80.77
Cloaked Images	70.65
Denoised Images	77.61

Figure 6: PCA for Feature Extractor ResNet



posed by the added noise. However, on applying the model to denoised images, there was a notable improvement, with accuracy rising to 77.61%. This improvement indicates that our denoising model, DnCNN, effectively mitigates the impact of cloaking. The table is shown in Table 1.

4.2 Final Results Analysis

We adopt two feature extractors, ResNet and AlexNet, to train our facial recognition models using three separate datasets: mixed, cloaked, and denoised.

4.2.1 ResNet

Consistent with our expectations, the model trained on cloaked images exhibits the lowest accuracy, both on original images and mixed images, shown in Table 2, highlighting the effectiveness of the cloaking process in disrupting facial recognition. In contrast, the model trained on denoised images performs better than the one trained with cloaked images, suggesting that denoising effectively counteracts some of the impacts of cloaking.

PCA for ResNet. In our Principal Component Analysis (PCA) on the ResNet feature extractor, we gain additional insights. As Figure 6 shows, the denoised images (represented by green spots) are closer to the original images (red spots) in the feature space, indicating successful denoising. There's a surprising minimal difference between the cloaked (blue spots) and original images, possibly due to our use of a basic-level Fawkes model and a limited perturbation budget. From these findings, we currently conclude that our denoising model can partially counteract the effects of the Fawkes model.

Table 2: Final Results on ResNet

Score on Feature Extractor ResNet			
Training Set for Training Facial Recognition Model	**Accuracy** for testing on **Original Images**	**Accuracy** for testing on **Mixed Images**	Feature Extractor
Crawled Images	93.11	92.20	ResNet
Cloaked Images	85.52	84.65	ResNet
Denoised Images	88.24	87.30	ResNet

Table 3: Final Results on AlexNet

Score on Feature Extractor AlexNet			
Training Set for Training Facial Recognition Model	**Accuracy** for testing on **Original Images**	**Accuracy** for testing on **Mixed Images**	Feature Extractor
Crawled Images	80.48	78.41	AlexNet
Cloaked Images	77.21	76.81	AlexNet
Denoised Images	80.44	78.56	AlexNet

Figure 7: Comparison Images

4.2.2 AlexNet

When we switch to using AlexNet as the feature extractor, similar patterns emerge. The facial recognition model trained with cloaked images shows the lowest accuracy. Yet, the model trained with denoised images often matches or even surpasses the performance of the one trained with images crawled from the internet. This indicates that our denoising model does more than just remove the perturbations introduced by Fawkes; it also appears to improve the model's overall ability to generalize (data in Table 3).

4.3 Comparison Images

As observed in Figure 7, Fawkes adds minor alterations to the original images, which are challenging to detect by the naked eye. Tiny variations are noticeable around the nose area. Both Fawkes model and denoising model do not significantly distort photos.

5 Limitations and Future Work

Looking ahead, our next steps involve conducting experiments with varying degrees of cloaking intensity in the Fawkes Model – low, medium, and high. This will allow us to deepen our understanding and enhance the robustness of facial recognition systems against these kinds of adversarial tactics.

6 Conclusion

In summary, our research demonstrates that denoising models, such as DnCNN, can markedly enhance the effectiveness of facial recognition systems, particularly in handling images affected by cloaking systems like Fawkes.

7 Statement of Work and Collaboration

Tong Zhou mainly focuses on the codes, and Xinbei Zhu is in charge of the creation of presentation slides and the final report. Our online team communication is done over Slack. We have an in-person meeting once a week to discuss approaches, assign work, and check our progress. We collaborate by having a team git repository.

References

[1] Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H. & Zhao, B.Y. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. arXiv:2002.08327 (2020)