Efficient Image Poisoning as Defense: Disrupting Profile Matching on OSNs and Preserving Human Comprehension

Kousar Kousar
Bilkent University
Ankara, Türkiye
kousar.kousar@bilkent.edu.tr

Aqsa Shabbir
Bilkent University
Ankara, Türkiye
aqsa.shabbir@bilkent.edu.tr

Ecem İlgün
Bilkent University
Ankara, Türkiye
ecemilgun97@gmail.com

Mehmet Kadri Gofralılar
Bilkent University
Ankara, Türkiye
kadri.gofralilar@bilkent.edu.tr

Noor Muhammad
Bilkent University
Ankara, Türkiye
noor.muhammad@bilkent.edu.tr

ABSTRACT

Abstract: Facial recognition systems are essential to modern security and surveillance, However, their widespread deployment has raised significant privacy concerns, necessitating effective obfuscation techniques to protect individual identities. Current methods, including adversarial face synthesis and clean-label poisoning attacks, demonstrate varying degrees of success in evading Automated Facial Recognition (AFR) systems. This paper introduces a novel approach combining the boundary box technique with salient feature analysis and Gaussian blur to obfuscate facial images effectively. Our methodology evolves around two different obfuscation methods. The first one applies Gaussian blur within a boundary box around the face. The second one identifies the most distinctive facial features —such as the eyes— using deep learning models, and reduces the pixel density to half selectively within a boundary box around these features. The same obfuscation is also tested on the whole face. These targeted obfuscations balance the privacy-utility trade-off, ensuring that the images remain recognizable to human observers while fooling AFR systems. Testing against multiple state-of-the-art face recognition models validates the robustness of our approaches. By optimizing the degree of blurring, we achieve effective privacy protection without compromising the usability of the images for other purposes, obtaining a 5% drop in accuracy with the first approach. However, the second approach unexpectedly resulted in an average increase of 1.5% in accuracy. Our code is publicly available on GitHub [1] and can be used for replicating our study. This research highlights the need for advanced obfuscation techniques to safeguard privacy in an era of pervasive facial recognition technology.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Face recognition, Privacy, Privacy protection, Facial obfuscation, Obfuscated images, Privacy concerns, Adversarial perturbations

ACM Reference Format:
Kousar Kousar, Aqsa Shabbir, Ecem İlgün, Mehmet Kadri Gofralılar, and Noor Muhammad. 2024. Efficient Image Poisoning as Defense: Disrupting Profile Matching on OSNs and Preserving Human Comprehension. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Facial recognition systems have become integral to modern security and surveillance infrastructures. These systems are employed in numerous applications ranging from unlocking mobile devices and authorizing transactions to enhancing security at airports. AFR systems leverage advanced deep learning models, particularly Convolutional Neural Networks (CNNs), to achieve remarkable accuracy in identifying and verifying individuals. These systems enhance accuracies exceeding 99% in True Accept Rate (TAR) at 0.1% False Accept Rate (FAR) [1], underscoring their effectiveness in controlled environments.

However, the applications of AFR systems have raised significant privacy concerns. As these systems become more widespread, the potential for misuse and unauthorized surveillance increases. Further, protecting individuals' privacy has become a critical challenge. This has developed a need for various obfuscation techniques designed to defeat AFR systems without compromising the utility of the images.

Obfuscation techniques are employed to modify images in a way that they can evade facial recognition systems while remaining recognizable to human observers. These methods introduce fine alterations that are often unnoticeable to the human eye but significantly impact the performance of machine learning models. The primary goal is to create adversarial examples that either cause misidentifying or prevent identification by the AFR system.

Current approaches to image obfuscation include the generation of adversarial perturbations, the application of geometric transformations, and the use of Generative Adversarial Networks (GANs).

[1] https://github.com/mehmetkadri/cs577-final-project

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

For instance, Deb et al. [2] utilize GANs to produce minimal perturbations in salient facial regions, effectively fooling state-of-the-art face matchers with high success rates, including white-box and black-box models. Similarly, Shafahi et al. [3] uses clean-label poisoning attacks to manipulate training datasets to inject adversarial examples, thereby compromising the integrity of the recognition model without directly altering test images. Such attacks are effective even without control over the labelling process. Techniques described by Raynal et al.[4] focus on hiding identifiable features while maintaining usability for other tasks. Methods include adding noise, blurring, and applying masks to critical facial regions. These changes make it harder for the facial recognition system to accurately identify people, but the images can still be used for other purposes where identification is not needed.

Our proposed approach leverages a combination of the boundary box technique with salient feature analysis and Gaussian blur to obfuscate facial images effectively. The boundary box technique uses deep learning models to detect the most salient features in a face image. These features are then marked within a boundary box, highlighting the regions to be targeted for obfuscation. The Gaussian blur is selectively applied to the regions within the boundary box. The intensity of the blur is carefully calibrated to ensure that the obfuscation is sufficient to fool AFR systems while keeping the image recognizable to humans. This selective blurring technique preserves the integrity of the image outside the boundary box, maintaining its overall quality and usability.

A critical aspect of any obfuscation technique is the privacy-utility trade-off. This trade-off refers to the balance between the degree of privacy protection provided by the obfuscation method and the usability of the resulting image for other purposes. In our proposed approach, this trade-off is carefully managed by targeting only the most salient features of the face for blurring while leaving the rest of the image intact. By focusing on the boundary box on the salient features, we can achieve obfuscation where it matters most for facial recognition systems while minimizing the impact on the overall visual quality of the image. The selective application of Gaussian blur within the boundary box allows for fine-tuned control over the obfuscation process, enabling a balance between privacy and utility. Users can adjust the level of blurring to meet their specific needs, achieving optimal protection without compromising too much on image usability.

In summary, our contributions are:

- Develop an effective method to obfuscate facial images to fool facial recognition systems while maintaining the usability of the images for other purposes.
- Use deep learning models to detect the most distinctive facial features. Mark these features within a boundary box for targeted obfuscation.
- Apply Gaussian blur selectively to the regions within the boundary box. Control the intensity of the blur to ensure effective obfuscation while maintaining the recognizability of the face for humans.

2 RELATED WORK

In this section, we outline the related work in the literature. First, we review the above-mentioned keywords and search venues we have used to access the related work. Next, the study selection criteria are discussed in more detail. To have a better view of the selected works in the literature, we provide a concise review of each paper.

2.1 Search Venues

We utilize a hierarchical search heuristic to find the most meaningful papers related to our work. First, we gather the most notable works that have addressed similar approaches to mitigate associated privacy risks. Since each of these papers provides a link to other related works in the domain, we use this chance to find more related works in the next step. We reiterate the same approach for the next level of the hierarchy until we reach satisfactory results both in terms of quality and quantity of the papers.

To search for these papers, we consider various platforms which are mentioned below:

- Google Scholar
- Institute of Electrical and Electronics Engineers Xplore (IEEE Xplore)
- Association for Computing Machinery Digital Library (ACM Digital Library)
- SpringerLink

Analyzing the related papers published on these platforms provided a broader exploration and understanding of the field.

2.2 Study Selection Criteria

Our selection criteria are focused on several aspects of the published papers. Initially, the papers are selected from the most recent articles published in the privacy domain. We limit our work to the published papers in the last 10 years.

The papers are also selected based on their relevance to the main topic. In this regard, the selected papers are classified into three major groups. Papers that focus on:

- Understanding of Face Obfuscation Techniques
- Attacking Facial Recognition Model
- Defeating Image Obfuscations

In the upcoming subsection, selected works are analyzed in detail.

2.3 Summaries of the Studies

We analyze eleven papers in total. Three of them will be mentioned in the "Understanding of Face Obfuscation Techniques" subsection, another four will be in the "Attacking Facial Recognition Model" subsection and the final four in the "Defeating Image Obfuscations" subsection.

2.3.1 ***Understanding of Face Obfuscation Techniques.*** Understanding of Face Obfuscation Techniques explores various methods used to anonymize facial images, ensuring privacy and security in digital media, we analyze below mentioned related paper that discusses this matter.

The study by Yang et al. [5] examines the impact of face obfuscation on the accuracy of visual recognition models trained on ImageNet, with a focus on minimal performance impact. Their methodology involves using simple obfuscation methods, such as blurring and overlaying, and analyzing their effects on model accuracy and feature transferability. Employing neural networks like AlexNet, VGG, and ResNet, among others, the research provides a comprehensive evaluation of the impact of face obfuscation on recognition accuracy across various tasks. This is captured through a well-structured methodology integrating a threat model, attack strategy, and empirical validation. However, their approach has a negative impact on the utility of images.

The study by Ren et al. [6] proposes a novel approach for learning to anonymize faces in videos using adversarial training, aimed at removing privacy-sensitive information while retaining spatial action detection capabilities. The paper introduces a method that contrasts with conventional anonymization techniques like masking, blurring, and noise addition, and highlights the benefits in action detection performance without compromising privacy. The methodology leverages an adversarial training framework that includes two competing models: a video anonymizer and a discriminator. The video anonymizer modifies original videos to anonymize faces, while the discriminator attempts to extract privacy-sensitive information from these videos, performing pixel-level modifications to minimize the impact on action recognition within the videos. The neural network models utilized include GANs for face anonymization, Faster R-CNN for spatial action detection, and SphereFace-20 for face recognition acting as the discriminator. Matrix evaluation involves metrics like mean Average Precision (mAP) for action detection accuracy and error rates in face verification, using datasets such as DALY and JHMDB to evaluate the performance in realistic scenarios. The study's strengths lie in its innovative adversarial approach that effectively balances privacy protection with high-quality action recognition, supported by empirical results and a user study that demonstrates superior performance over traditional techniques.

The study by Kumar et al. [7] explores novel methods for face verification by utilizing attribute and simile classifiers, which significantly enhance accuracy without requiring image alignment. Their study utilizes a dataset, PubFig, along with LFW to demonstrate its effectiveness under various conditions, including different poses, illumination, and expressions. The methodology introduced comprises two key components. Attribute Classifiers utilize 65 describable visual traits (such as gender, race, and age) and employ binary classifiers to verify faces based on the presence or absence of these traits. Simile Classifiers assess the similarity between face parts and a set of reference individuals, using binary classifiers that eliminate manual labelling required by the attribute method. The study uses support vector machines (SVMs) with radial basis function (RBF) kernels for classification. The performance metrics detailed in the paper show significant error rate reductions of 23.92% and 26.34% for the attribute and simile classifiers, respectively, with a combined improvement of 31.68%. The study demonstrates significant improvement over previous models in face verification accuracy.

2.3.2 ***Attacking Facial Recognition Model***.

Attacking facial recognition models involves techniques designed to deceive detection by systems that identify individuals based on their facial features. Such studies are vital for understanding potential vulnerabilities in facial recognition technology and developing more secure and reliable systems. We analyze below mentioned related paper that discusses this matter.

The study by Shafahi et al. [3] explores a stealthy data poisoning attack on neural networks. This method involves introducing specially crafted, correctly labelled instances into the training set to manipulate model behaviour during testing. The research utilizes an optimization approach to craft poison instances that closely match target instances in the feature space while maintaining their correct label in the input space. The methodology uses neural networks, specifically modified versions of InceptionV3 and AlexNet, tailored to datasets like ImageNet and CIFAR-10. The matrix evaluation includes measures such as the Attack Success Rate, which measures the percentage of instances where the poisoned model misclassifies the target instance as intended by the attacker, and the Model Accuracy Impact, which observes the change in overall model accuracy due to the introduction of poison instances. Strengths of the study include the technique's ability not to require altering the labels of training instances, thus providing a subtle approach to traditional poisoning attacks. However, the complexity of crafting effective poison instances that require in-depth knowledge of the model's feature space presents a significant challenge. This method manipulates neural network behaviours, offering insights into the vulnerabilities of machine learning systems.

The study by Sabour et al. [8] explores how precise adversarial images can significantly alter deep neural network (DNN) representations, making an image's internal representation mimic that of a different one. The paper highlights the unexpected vulnerabilities in how DNNs interpret and process visual information. The methodology employed involves creating adversarial images where the internal DNN representation of one image is altered to closely resemble that of another, using minimal visible changes. This is achieved through gradient-based optimization to minimize the Euclidean distance between the DNN representations of the adversarial and target images, under constraints that keep the visual perturbations imperceptible. The models utilized in this study include the BVLC Caffe Reference model (Caffenet), with additional testing on networks like AlexNet, GoogleNet, and VGG CNN-S to validate the generality of the findings across different architectures. The evaluation of the success of adversarial images is based on how closely their internal representations match those of target images across various DNN layers, including analyses of Euclidean distances, nearest neighbour distances, and rank statistics in feature space. Strengths of this research include introducing a new perspective on the potential for adversarial images to manipulate internal DNN representations beyond mere misclassifications, demonstrating the manipulation across multiple network architectures, and enhancing the understanding of DNN vulnerabilities which could guide future network architecture and defence mechanism developments.

The study by Chandrasekaran et al. [9] is aimed at enhancing privacy by adding strategic perturbations to user images, making them unrecognizable to commercial face recognition services while maintaining their appearance to human viewers. This research employs datasets such as LFW, VGGFace2, and Celeb. The methodology addresses the challenges posed by the black-box nature of these services and the limitations of current adversarial attack literature on metric networks. It involves designing novel loss functions and leveraging transferability to ensure the effectiveness of the perturbations against unknown models. Neural network models utilized include FaceNet, employing the triplet loss architecture, and CenterNet, using the center loss architecture. The evaluation metrics for the study focus on the top-1 matching accuracy for both targeted and untargeted attacks. For targeted attacks, the success metric is the ratio of successful matches to the intended target label over the total number of attacks. For untargeted attacks, it is the ratio of cases where the label of the adversarial example is not the true (source) label, indicating the effectiveness of the attack. The strength of the study lies in demonstrating the manipulation across multiple network architectures. However, a significant challenge highlighted is the balance between making perturbations effective at avoiding detection and keeping them unnoticeable to human viewers, which underscores the complexity of designing adversarial examples that are both effective and subtle.

The study by Deb et al. [2] focuses on crafting adversarial images designed to deceive facial recognition systems. Utilizing the LFW dataset, this study emphasizes creating images that both possess high perceptual quality and can be generated rapidly. The major aim is to achieve high evasion success rates against black-box face matchers, presenting a challenge to current face recognition technologies. The methodology leverages GANs to generate minimal perturbations in crucial facial regions, resulting in adversarial images capable of misleading face recognition systems effectively. The evaluation of these adversarial images is conducted using Structural Similarity (SSIM), assessing the visual similarity between the original and the adversarial images to ensure that the perturbations are not perceptible to humans. This approach demonstrates a high success rate in evading detection by black-box face recognition systems while maintaining the images' high perceptual quality.

2.3.3 *Defeating Image Obfuscations*.

Defeating image obfuscations focuses on methods to bypass techniques that hide visual content to prevent machine recognition. This area of study is essential for enhancing image analysis technologies and improving the accuracy of systems in environments where images are manipulated. We analyze below mentioned related paper that discusses this matter.

The study by Garofalo et al. [10] explores the impact of adversarial machine learning on face authentication systems by demonstrating a poisoning attack that significantly compromises system integrity. This research is the first to deploy a poisoning attack against an authentication system based on a state-of-the-art face recognition technique. The methodology involves a detailed threat model, the development of an attack strategy, and empirical testing using the OpenFace framework combined with an SVM classifier. The attack strategy is to inject a single crafted adversarial image into the training data, which poisons the SVM classifier. The study employs the FaceNet CNN model, which is designed to extract identifying features from faces to generate templates, subsequently classified using SVM. It employs a Triplet Loss function during training to optimize the distance metrics for face recognition. The primary strength of the research lies in addressing a relevant issue focusing on the robustness of face recognition systems against adversarial attacks.

The study by Li et al. [11] focuses on anonymizing faces in images while maintaining the ability to recognize identities using facial recognition systems. Their research addresses the contradiction observed in conventional face obfuscation techniques by preserving identity features while anonymizing other facial attributes. The methodology integrates two main components: identity-aware region discovery and identity-aware face confusion. These elements work together to obfuscate identity-independent attributes while retaining identity-relevant features, thus ensuring the face remains recognizable by facial recognition technologies. The research employs a conditional generative adversarial network (cGAN) approach using StarGAN as the base generation network for the face obfuscation model. For evaluating the effectiveness of the face anonymization, the study uses metrics like LPIPS Distance to measure the perceptual similarity between the generated and original images, focusing on weighted distances between deep features of images. It also uses FID Distance to assess the divergence between the distributions of real and generated images, where a lower FID score indicates higher quality and diversity, making the generated images closer to the real ones. This approach applies an advanced technique to retain recognizable features while effectively anonymizing other aspects of faces. The study's main challenge lies in the complexity of attributes that must be successfully obfuscated without affecting identity recognition.

The study by McPherson et al. [12] investigates the limitations of obfuscation techniques when faced with advanced neural network models. They specifically focus on artificial neural networks' potential to recover hidden information from obfuscated images. This study underscores the increasing need to reassess traditional privacy-preserving methods under the lens of modern deep-learning capabilities. The methodology utilizes artificial neural networks to attack obfuscated images, demonstrating that even heavily obfuscated images can be decoded to reveal sensitive information. The research uses deep convolutional neural networks with dropout regularization to enhance their robustness and effectiveness in breaking through obfuscations. The study employs accuracy, precision, recall, and F1 scores to compare the network's ability to correctly identify and reconstruct information from obfuscated images against a baseline. This provides a comprehensive analysis of the vulnerability of image obfuscation techniques to neural network-based attacks. The major strength of this paper lies in providing a detailed deficiency in current image obfuscation techniques.

The study by Raynal et al. [4] focuses on defending against facial recognition by developing and evaluating obfuscation techniques to preserve privacy in machine learning applications. The goal is to prevent the identification of individuals in images, ensuring that identifiable features such as faces are obscured to protect privacy. The proposed obfuscation methods include mixing images, pixel grafting, shuffling pixels, adding noise, pixelizing, and blurring, all of which are designed to conceal identities without significantly degrading the accuracy of machine learning models trained on these images. The authors validate their approach through user surveys and tests with AI-based recognition systems, demonstrating that their techniques effectively obscure personal identities while maintaining the utility of the obfuscated images for training purposes.

3 SYSTEM OVERVIEW

3.1 Problem Statement

The increasing integration of facial recognition systems into security, surveillance, and personal device authentication has raised significant privacy concerns. Despite their high accuracy and efficiency, these systems pose substantial risks of unauthorized surveillance and identity tracking. The primary challenge is to develop effective image obfuscation techniques that protect individual privacy without compromising the usability and recognizability of images for non-identification purposes. Our project aims to address this issue by balancing the need for privacy protection with the requirement to maintain the practical utility of the images.

3.2 System and Threat Model

System Model: Our system model comprises several key components to achieve effective image obfuscation. First, we obtained facial images from the LFW dataset [13] and the FaceScrub dataset [14]. Then for our first approach, we apply Gaussian Blur on the whole face in images obtained from the LFW dataset for obfuscation. This is done with the intensity of the blur carefully calibrated to ensure that the obfuscation is sufficient to fool AFR systems while keeping the image recognizable to humans. For the second approach, we employ a deep learning model to identify and mark the most salient facial features such as eyes. These features are enclosed within boundary boxes, which describe the regions targetted for obfuscation. This obfuscation approach involves reducing the pixel density selectively within these boundary boxes, making them darker. This obfuscation is also applied to the whole face for comparing results in further steps.

Threat Model: Our threat model assumes that adversaries have access to advanced AFR systems capable of high-accuracy identification and verification. These adversaries can obtain and analyze facial images from various public and private sources, including databases, social media, and surveillance footage. Their primary goals are to identify individuals in obfuscated images and bypass privacy protection measures to track individuals without their consent. To defend against these threats, our system focuses on the most unique facial features and applies Gaussian blur to these regions, interrupting the AFR system feature extraction process. Testing against a state-of-the-art AFR system ensures the effectiveness of our obfuscation method, providing privacy protection while maintaining the utility of images.

3.3 Methodology

In this section, we explain our methodology and approaches in detail. First, we share the difficulties and problems we encountered while trying to replicate the study Face-Off[9]. Next, we will explain the two approaches we implemented in this research and present results for both.

3.3.1 *Replication of Face-Off[9].*

At first, we tried to replicate the Face-Off study from their GitHub repository [2]. Due to the study being a few years old, and the authors not using the latest libraries even for the time of the development of their method, the exact environment was not available. Some libraries were no longer downloadable in the versions used, especially Tensorflow 1.9 and Python 3.5.2 were no longer supported. Eventually, we were able to set up an environment with those versions, however, then we had problems setting up cleverhans library. These two libraries and Python were crucial for the Face-Off code to run.

It took us 2 weeks to get the right combinations but in the end, we were able to set up an environment with Python 3.7 and Tensorflow 1.13.1, without explicitly changing the other requirements. But even so, the same configurations didn't work for some of us and kept getting errors. With these versions, we were also able to install cleverhans. Then we had to start converting the original code to support the later versions of the libraries. Many changes were necessitated by differences in function names, parameter lists, and library structures between the versions originally used and those we had access to. To our surprise, we ended up having to change over 300 lines in total. Each of these lines was found out when we got another error trying to run the Carlini-Wagner targeted attack [15] module of the Face-Off repository. Each time we got the error deeper into the pipeline and the new errors we encountered started to be very costly. Unfortunately, having to run the attack each time to find where the attack would succeed or give an error took 3 weeks in total.

Despite these modifications, we managed to replicate the core functionality of the Face-Off study. However, we faced an additional challenge when running the Carlini-Wagner attack. For reference, the Carlini-Wagner attack is a well-known adversarial attack on neural networks. The optimization problem they formulated is given by:

$$\begin{aligned} &\min \|\delta\|_p \\ \text{s.t.}\quad &\|\delta\|_p \le \epsilon, \\ &G(x+\delta, t) \le 0 \end{aligned} \tag{1}$$

where δ is the perturbation, p represents the norm, ϵ is a small constant, G is a function measuring the distance between the modified input and the target class t, and x is the original input image. The G measure in particular to Face-Off is Hinge Loss defined in equation (3), where $z = x + \delta$ is the perturbed sample.

To define the Hinge Loss properly, we first define d_f, the distance it uses. d_f is defined by a deep embedding f_θ and a set A_y. The deep embedding in question is a function from sample space $\mathbf{X}$ to $\mathbb{R}^m$, where θ is a parameter chosen from the parameter space. The set A_y is the subset of samples with label y from the whole dataset in use.

[2] https://github.com/wi-pi/face-off

With a particular f_θ and A_y, d_f becomes:

$$d_f(z, A_y) = \|f_\theta(z), f_\theta(A_y)\|_2 = \frac{1}{|A_y|} \sum_{(a_i, y) \in A_y} d_f(x, a_i) \quad (2)$$

where y is the perturbed sample z's true label. With this d_f, we can define Hinge Loss G as follows:

$$G(x + \delta, t) = \left[d_f(x + \delta, A_t) - \max_{y \neq t} d_f(x + \delta, A_y) + \kappa \right]_+ \quad (3)$$

where the margin κ denotes the desired separation from the source label's samples. Further details of the deep embedding function and more are explained in detail in Face-Off.

In the end, we encountered a "'NoneType' object is not subscriptable" error. This error occurred because δ in the Carlini-Wagner attack was not assigned at all.

The root cause of this problem was elusive, as it happened inconsistently across different κ and δ combinations for different target individuals. Some target individuals experienced no errors, while others did. We suspect this inconsistency is due to changes in a library from the time the method was originally developed to now. However, this issue could not be resolved by a simple correction from the previous syntax to the new syntax. We believe the implementation of a function that Face-Off called during their implementation of the Carlini-Wagner attack has changed. We narrowed down the possible function to `attack_batch_l2` in their implementation of Carlini-Wagner, and possibly in `sess.run` function called in `attack_batch_l2`, where sess is the tensorflow session. Our results and the 'Nonetype' error we encountered are replicable through Github [3]. Given our limited time, we decided to implement our own approach instead of attempting to fix the Face-Off implementation to ensure we had some results.

3.3.2 ***Boundary Box Obfuscation with Gaussian Blur Applied***. In this method, we focus on selecting specific regions in the image and changing them to fool the model. We decided to focus on the face region and applied Gaussian blur to it.

An LFW dataset[13] called deepfunneled is used for this method. The dataset contains images of several famous people around the globe including actors, politicians, athletes, etc. The dataset originally contains more than thirteen thousand images. However, due to constraints in time and processing capabilities, we only picked a small sample of 618 images from that dataset to evaluate the potential of this methodology.

Due to a lack of access to the industry-used recognition models, surrogate models were used to test the efficiency of our methodology. One such state-of-the-art model is FaceNet. The model uses a deep convolutional neural network to learn face mapping and then the similarity between faces is calculated using squared Euclidean distance. It uses a triplet loss function as a loss function.

As FaceNet only accepts input images of the *160x160*, so the sampled dataset was processed and resized to the desired state. The model was trained on normal boundary-boxed 618 images and accuracy was tested between the original images and the Gaussian blurred images. Example images are given in Figure 1.

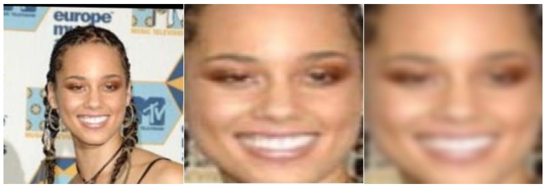


Figure 1: Picture of a celebrity, boundary boxed and Gaussian blurred

3.3.3 ***Boundary Box Obfuscation with a Focus on Salient Features***. In this method, we focus on obfuscating only the salient features (e.g. eyes) in images. The obfuscation we applied was to decrease the density of the pixels inside the boundary boxes of the salient features, making them darker.

The dataset used for this approach was the FaceScrub dataset [14]. This dataset consists of two files; "facescrub_actors" and "facescrub_actresses". The files included features such as "name", "image_url" and "boundary_box_borders" for 530 unique identities and a total of 106,863 samples. Initially, we organized these files by merging them and changing the format into "name", "img_id", "url" and "face_location". Due to the dataset size being too much, we had to take a small sample from it. We chose 3 as the number of images per person to be downloaded and 200 as the number of person to apply obfuscations on. These can be changed by providing respective parameters in the replication package. Then, 3 images per person are downloaded and 200 random people are chosen among them and saved for stability. Afterwards, obfuscation is applied to all 3 downloaded images for every randomly chosen person as follows:

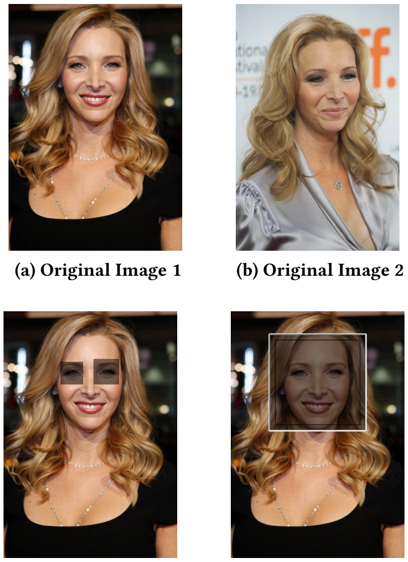
1. First, the salient features (i.e. eyes) are found by the utilization of *.CascadeClassifier()* function from OpenCV, which is a Haar feature-based cascade classifier with the parameter "haarcascade_eye.xml".
2. After getting the boundaries of the salient features, the pixel density is decreased to half inside these boundaries. This operation makes the areas darker.
3. The resulting image is saved to a new directory under the person's name with the image ID.

Afterwards, we wanted to apply this method for the whole faces as well, to compare results. For that reason, the obfuscation is applied to the same original images one more time. However, instead of applying the obfuscation on the salient features only, it is applied to the whole face using the "face_location" information provided by the dataset. Example obfuscated images obtained using this approach are given in Figure 2.

For simplicity, the images that are used and obtained from every person will be referred to as follows:

- **Original_1** : The first original image of the person that the obfuscations are applied later on such as Figure 2.a.
- **Original_2** : A second original image of the same person such as Figure 2.b.
- **Salient_Obfuscated** : The version of the first original image which salient feature obfuscation applied such as Figure 2.c.
- **Face_Obfuscated** : The version of the first original image which whole face obfuscation applied such as Figure 2.d.

[3] https://github.com/theyusko/face-off



(a) Original Image 1

(b) Original Image 2

(c) Salient Features Obfuscated Image

(d) Whole Face Obfuscated Image

Figure 2: Resulting images after obfuscation

4 RESULTS

The first methodology is tested by comparing the accuracy of the trained surrogate model to detect the original images that it was trained on with the Gaussian blurred version of those same images. Boundary box Gaussian blur actually showed some promise as it heavily decreased the model's accuracy.

Total images processed	608
Correctly predicted images(without blur)	443
Correctly predicted images(with blur)	411
Accuracy on images (without blur)	72.86%
Accuracy on images (with blur)	67.59%

Table 1: Accuracy results for Gaussian blur boundary box

The 5% approximate decrease in accuracy shows the potential of the boundary box Gaussian blur in evading recognition models. However, this result is still not very reliable due to the small sample size used for training and testing the model.

The second methodology is tested by the following system for all 200 randomly chosen person:

(1) Match accuracy between **Original_1** (refer to 3.3.3) and **Original_2** (refer to 3.3.3).
(2) Match accuracy between **Original_1** and **Salient_Obfuscated** (refer to 3.3.3).
(3) Match accuracy between **Original_1** and **Face_Obfuscated** (refer to 3.3.3).
(4) Match accuracy between **Salient_Obfuscated** and **Face_Obfuscated**.

The final results for this methodology are given in Figure 3.

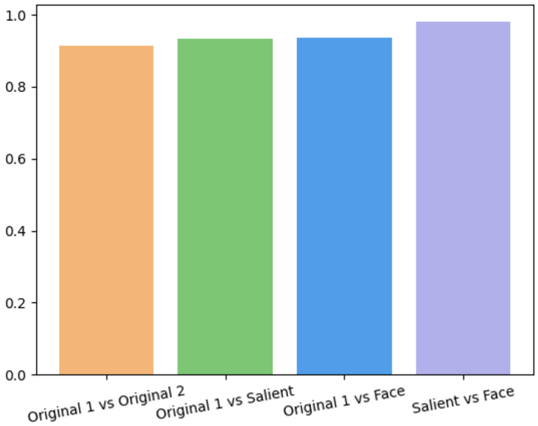


Figure 3: Accuracy results for Boundary Box Salient Feature Obfuscation

In the bar chart given above, we can see that our second methodology did not obtain good results. The first bar demonstrates the accuracy between two different images of the same person, the second demonstrates the accuracy between an original image and the salient features obfuscated version, the third demonstrates the accuracy between an original image and the whole face obfuscated version, and the final bar demonstrates the accuracy between two different obfuscated versions of the same image. These results can be interpreted as follows:

- Our proposed salient feature obfuscation increased the accuracy, which means the matching rate is higher and the accuracy is lower compared to no obfuscation.
- Our second proposed whole-face obfuscation also increased the accuracy even more, which means the matching rate is higher and the accuracy is lower compared to no obfuscation.
- When our proposed obfuscation methods are applied to the same image and they are tried to match, the accuracy is at its highest. This means that these two methods are providing quite similar results. Therefore, omitting one could be beneficial for future studies.

5 FUTURE WORK

Future work will commence by replicating the methods from the [9] by implementing its adversarial attacks and defences and creating a public code repository for transparency. Enhancements to our obfuscation techniques will involve refining the boundary box method, Gaussian Blur application, and making changes in the Salient Feature Obfuscation approach, integrating advanced machine learning techniques for practical applicability. A comprehensive analysis will follow, detailing success rates and utility impacts, ensuring privacy protection against unauthorized facial recognition systems while preserving the utility (i.e. the image quality).

6 CONCLUSION

With facial recognition systems being integrated into most parts of daily life and developed more and more every day, security systems have been greatly enhanced. However, the enhancement of these technologies has also raised privacy concerns regarding the protection of individual identities through OSNs (Online Social Networks). These concerns increased the studies to create a defence mechanism utilizing obfuscations against image matching, especially against profile matches in OSNs. Considering the nature of this type of attack, achieving this was a hard task. However, the previous studies managed to obtain decreased accuracies in matches. Even though that is the case, the resulting images in most of these studies were too corrupted to share in OSNs. One particular study, Face-Off[9], managed to obtain good results. The images were not corrupted for the human eye, but the deep networks were confused and the accuracy dropped significantly. So, we decided to try and replicate the study done in Face-Off and make alterations to their methodology. Upon trying to replicate Face-Off, we encountered issues such as errors that are almost impossible to debug in a short period of time. Therefore, we decided to come up with a novel methodology. After going through the previous studies done in this domain and other relevant topics, we decided to focus on the Boundary Box Obfuscation method. First, we tried boundary box obfuscation by applying Gaussian blur. The results were promising since there was a 5.27% decrease in the accuracy. However, considering the dataset was too small for a general conclusion, we decided to enlarge our dataset to reach a more generalized conclusion. Then, the boundary box obfuscation is tried with a focus on salient features. However, the results were not satisfying, since the accuracy is even higher compared to making no obfuscation, therefore needs improvement in the future.

REFERENCES

[1] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on evaluation of 2d still-image face recognition algorithms, nistir 7709 [r]," *Maryland Gaithersburg: National Institute of Standards and Technology*, vol. 8, 2011.

[2] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB).* IEEE, 2020, pp. 1–10.

[3] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[4] M. Raynal, R. Achanta, and M. Humbert, "Image obfuscation for privacy-preserving machine learning," *arXiv preprint arXiv:2010.10139*, 2020.

[5] K. Yang, J. H. Yau, L. Fei-Fei, J. Deng, and O. Russakovsky, "A study of face obfuscation in imagenet," in *International Conference on Machine Learning.* PMLR, 2022, pp. 25 313–25 330.

[6] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 620–636.

[7] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th international conference on computer vision.* IEEE, 2009, pp. 365–372.

[8] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," *arXiv preprint arXiv:1511.05122*, 2015.

[9] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, "Face-off: Adversarial face obfuscation," *arXiv preprint arXiv:2003.08861*, 2020.

[10] G. Garofalo, V. Rimmer, D. Preuveneers, W. Joosen *et al.*, "Fishy faces: Crafting adversarial images to poison face authentication," in *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.

[11] J. Li, L. Han, R. Chen, H. Zhang, B. Han, L. Wang, and X. Cao, "Identity-preserving face anonymization via adaptively facial attributes obfuscation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3891–3899.

[12] M. Richard, S. Reza, and S. Vitaly, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[13] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[14] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 343–347, 2014.

[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.