Diploma Thesis

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Face anonymizer

Bc. Jiří Moravčík

Supervisor: Ing. Vojtěch Franc, Ph.D.
Study program: Open Informatics
Specialisation: Artificial Intelligence
May 2023

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Moravčík Jiří** Personal ID number: **483741**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Artificial Intelligence**

II. Master’s thesis details

Master’s thesis title in English:

Face anonymizer

Master’s thesis title in Czech:

Anonymizátor tváří

Guidelines:

The main goal of the thesis is to design and develop a framework for automated evaluation of face anonymization methods. The framework will apply a given face anonymizer on a set of images and evaluate quality of the produced result by well-chosen criteria that have a clear interpretation and allow one to judge about the use of the anonymizer in a particular application. The framework will be used to rank existing open source anonymizers based on the objective critera. The secondary goal of the thesis is to identify weaknesses of some of an existing method and to propose and implement its improvement.

Bibliography / sources:

- Hukkelas et al. DeepPrivacy2: Towards Realistic Full-Body Anonymization. WACV 2023.
- Klomp et al. Safe Fakes: Evaluating Face Anonymizers for Face Detectors. CVPR 2021.
- Maximov et al. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. CVPR 2020.
- Hukkelas et al. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. CVPR 2019.

Name and workplace of master’s thesis supervisor:

Ing. Vojtěch Franc, Ph.D. Machine Learning FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **06.02.2023** Deadline for master's thesis submission: **26.05.2023**

Assignment valid until: **22.09.2024**

Ing. Vojtěch Franc, Ph.D. Supervisor's signature	Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt	Student’s signature

Acknowledgements

I would like to thank my supervisor Ing. Vojtěch Franc, Ph.D. for his help and guidance with this thesis.

Furthermore, I would like to thank everyone that supported me while I worked on this thesis.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, 26. May 2023

Abstract

In this thesis, we investigate the field of face anonymization. The main contribution is a novel suite for benchmarking of face anonymization methods. Futhermore, we develop a framework that converts the task of face anonymization to the task of face swapping. We demonstrate that our framework is competitive to existing methods and strongly improves upon biases in facial attributes. Lastly, we employ our proposed benchmark suite to compare and evaluate several existing anonymization methods. We also perform a study by human annotators that evaluates photorealism of anonymization methods. The results show that even the best methods are not perfect in every criterion and there is a lot of room for future research.

Keywords: face anonymization, generative models

Supervisor: Ing. Vojtěch Franc, Ph.D.

Abstrakt

V této práci se zabýváme oblastí anonymizace tváří. Hlavním přínosem je nová srovnávací sada pro metody na anonymizaci tváří. Dále vyvineme metodu, která převádí úlohu anonymizace tváře na úlohu výměny tváří. Ukážeme, že naše metoda je konkurenceschopná vůči existujícím metodám a výrazně zlepšuje výsledky v rámci zaujatosti u atributů obličeje. Nakonec použijeme naši navrhovanou srovnávací sadu k porovnání a vyhodnocení několika existujících anonymizačních metod. Provedeme také studii s lidskými anotátory, která hodnotí fotorealismus anonymizačních metod. Výsledky ukazují, že ani ty nejlepší metody nejsou dokonalé ve všech kritériích a existuje mnoho prostoru pro budoucí výzkum.

Klíčová slova: anonymizace tváří, generativní modely

Překlad názvu: Anonymizátor tváří

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Anonymization

Anonymization is the process of removing or obfuscating personal information from arbitrary data. The main goal of anonymization is to protect the identity of individuals present in the data, while still preserving the usefulness of the data for common use cases such as analysis, research, or training of machine learning models. Some authors use the term *de-identification* instead of *anonymization*, in this work, we will prefer the latter.

Nowadays, governments impose strict regulations on data collection and analysis in hopes of preventing misuse of personal information. This is heavily supported by ever-growing amounts of personal data that are stored across many institutions and companies. One common example of such regulation is the General Data Protection Regulation (GDPR) in EU.

With that in mind, anonymization becomes essential in many fields such as healthcare, government services, law enforcement, or finance. These fields collect large amounts of personal data. Data sets can be of immense value to researchers and analysts; however, they would not be able to use them if the data were not anonymized.

1.2 Image and video anonymization

Image and video anonymization has recently gained a lot of attention. Image and video collection is common in today's society, and the amount of collected data keeps growing every year.

Collected images and videos are commonly used to train and evaluate machine learning models used for, e.g. object detection, image segmentation,

face detection. With the rise of deep neural networks, large datasets are needed to train the networks. The collected data can also be publicly displayed; in such cases, getting approval from all individuals present in the data is nearly impossible.

Given the existing legislation and regulations imposed on data and the need for large datasets, a conflict emerges between large-scale data collection and the need for data protection. Fortunately, there are methods that can anonymize images so that the data protection laws are not violated.

Traditional anonymization approaches such as blurring, image masking, or pixelization are widely used today. One example is Google StreetView; the faces and license plates are blurred, so they are not identifiable.

While such methods can protect the private data, they distort, or alter the images or videos in unnatural ways. Such distortions can make the data unusable in various domains.

One example is training computer vision models using the anonymized data. The models will be trained on data that do not correspond to the real data distribution and will perform poorly when deployed to production.

Another example is the perception of anonymized images or videos by humans. They will notice that images or videos are deformed by traditional anonymization methods, e.g. blurred faces or license plates. This can lead to decreased user-experience and reduction of believability in products using such anonymized images or videos.

Luckily, recent advances in generative models enable us to protect the private data without causing noticeable damage to the data.

1.3 Face anonymization

In this thesis, we restrict ourselves to face anonymization of images (photos). Generally, face anonymization refers to the process of concealing an individual's identity by altering a face in a photo or a video in such a way that the original identity is not easily recognizable. At the same time, all other visual attributes of the image are preserved. This can be done using generative techniques. We show an example of face anonymization in Figure 1.1.

With the rise of generative techniques in deep learning, impressive results have been achieved in the area of face generation. Specifically Generative Adversarial Networks (GANs), and recently Diffusion models (DMs) have been used to generate photorealistic faces.

While unethical use-cases like deep fakes create a lot of controversy in the area of generative models, there are use-cases that prove that generative

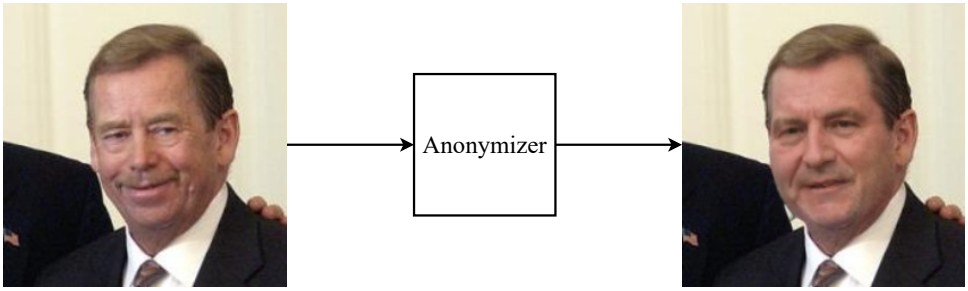


Figure 1.1: An example of face anonymization. The used anonymization method is DeepPrivacy2 [11].

modeling can have positive impact as well.

We describe the task of face anonymization as follows (note that this is one of many possible definitions):

1. The input is an image with a face.
2. The method generates a new face that should preserve original facial attributes but have a different identity.
3. Moreover, the generated face should be as photorealistic as possible.
4. All other visual attributes of the image are preserved.
5. The output is an image with the generated face.

Generative approaches to anonymization use deep learning models to replace private parts of the image with generated alternatives. Although there has been much progress over the past few years, generative anonymization remains a challenging task.

This is because it is difficult to generate faces that fit well into the original part of the image. This is due to many factors, such as the pose of the original face, hair, lighting conditions, etc. We review existing approaches in Chapter 2.

Lastly, we summarize our contributions as follows:

- We design and implement a benchmark suite, AnonyBench, consisting of five independent benchmarks that evaluate different aspects of face anonymization methods.
- We propose a general framework that converts the task of face anonymization to the task of face swapping.
- We use the proposed benchmark to evaluate representative examples of existing anonymization methods.

Each of the contributions will be thoroughly described in the following chapters. An outline of this work follows.

Outline of the Thesis

We begin by introducing the general concept of anonymization and discussing its usages. Then, we restrict ourselves to the task of face anonymization. We also specify the face anonymization task and describe the contributions of this thesis in Chapter 1.

Face anonymization has been widely studied in many research fields. We provide a review of the literature on face anonymization and related topics in Chapter 2.

The main contribution of our work is the benchmark suite. We describe the used metrics, philosophy behind them and give a technical overview over them in Chapter 3. We also describe the codebase in Appendix B.

We describe how to use a face swapping model to perform face anonymization in Chapter 4. This is a general framework that can be used with any viable face-swap model.

The evaluation of existing methods follows in Chapter 5. We evaluate several existing methods using our proposed suite. We also conducted a survey of the anonymized images by human evaluators. They were given a set of real and generated images and were supposed to classify, whether an image is real or generated.

We conclude the thesis and propose some ideas for future work in Chapter 6.

Chapter 2

Literature review

2.1 Face anonymization

There have been many attempts to create photorealistic face anonymization methods. Nevertheless, before the era of deep generative models, the field was dominated by traditional methods like black boxing, blurring or pixelization.

In spite of that, several attempts have been made to create anonymization methods using non-deep learning approaches. K-Same [30] anonymizes a given face with an average of k closest faces based on a distance metric. The average is taken either over the original pixels (k-Same-Pixel) or eigenvectors representing the images (k-Same-Eigen). Authors of k-Same-M [8] propose an extension of K-Same using Active Appearance Models (AAMs). [42] extends the AAM approach using different poses for the image database, specifically, the authors group the images into 3 groups: frontal view, look left and look right. For a more thorough review of the legacy anonymization methods, see [37].

With the rise of generative deep learning, models based on artifical neural networks have taken over as the state of the art.

Early attempts to use generative models include [27], where the authors propose to use a Generative Adversarial Network (GAN) [7] to synthesize a new face. The input of the GAN is a vector that encodes features about k-closest identities to the anonymized face. The features are extracted using a pretrained VGG face network. The final image replacement is done using facial landmarks and RANSAC for perspective transformation, blending is finalized using a Gaussian kernel mask. This approach is heavily inspired by the K-Same family of algorithms [30].

Authors of [36] propose a new approach to face anonymization using an adversarial training setting. They jointly train a generator that anonymizes the original image to remove privacy-sensitive information and a discriminator

that tries to extract privacy-sensitive information from the processed images. This is naturally an extension of the GAN formulation.

A novel approach to anonymization using head inpainting was introduced in [47]. The used approach splits the task into two subtasks: (1) facial landmark generation (2) head inpainting conditioned on facial landmarks. We note that inpainting methods do not explicitly control the identity of the generated face. The generative inpainting model is a GAN. U-Net [39] architecture is used for the head generator, the architecture of the GAN discriminator copies DCGAN [35]. [48] enriches the model with a parametric face decoder, the inpainting architecture stays the same.

Privacy-Protective-GAN (PP-GAN) [54] extends the GAN objective with two more criteria. First criterion measures verification loss for the original and generated faces. This ensures that the generated face holds an enough distance from the original face in the identity-related space. Second criterion measures SSIM [53] to make the generated face consistent in terms of luminance, constract, etc.

CIAGAN [26] uses a conditional GAN conditioned on original face landmarks to generate a new face, moreover, the new identity is chosen from a pretrained set of identities. On the other hand, CLEANIR [3] uses a Variational Auto-Encoder (VAE) instead of a GAN. The authors split the output vector into identity and facial attributes and then modify the identity only. This approach keeps the original facial attributes. CLEANIR still suffers from blurry quality given by the vanilla VAE architecture.

DeepPrivacy [12] and DeepPrivacy2 [11] use a GAN to inpaint new faces, specifically a U-Net [39] generator is used in both cases. DeepPrivacy uses original face landmarks when generating the new face, while DeepPrivacy2 does not. DeepPrivacy2 also introduces models for full-body anonymization.

CFA-Net [25] starts with a model that decouples the identity of the face from the rest of the image contents in the latent space. Then another model that modifies the original identity is used. The actual architecture is a GAN. AnonyGAN [4] takes the conditioning identity as another input, this gives more control over the identity to the user. The face generation uses landmarks that were aggregated from the original and conditioning identity by an attention mechanism allowing the generator to find the most suitable combination.

Recently, researchers have proposed a face anonymization model based on diffusion probabilistic models [46]. LDFA [18] uses the inpainting weights of Stable Diffusion [38] combined with the Euler Ancestral sampler to anonymize the datasets used in the context of intelligent transportation systems. Therefore, it is critical that the perception models for this use case are trained on data that contain realistic faces of pedestrians and cyclists. The authors used the model without any prompt. Note that the inpainting architecture of

Stable Diffusion follows the ideas from the LaMa [49] inpainting model.

Safe fakes [19] deserves a honorable mention, because it specifically focuses on evaluation of face anonymizers. This work attempts to evaluate anonymization methods with respect to face detection. The authors anonymize the WIDER FACE [57] dataset and train a face detector using the anonymized dataset. Then they measure the performance of the trained detectors using the original validation set. The authors report that DeepPrivacy [12] achieved the best performance.

2.2 Face swapping

While this work is concerned with face anonymization, we show that face anonymization can be solved using face swap models in Chapter 4.

Pre-deep learning face swapping approaches use landmarks, blending, and 3D models to replace the original face with a different one. [2] uses a three-stage pipeline that aligns the replacement face with the pose of the original face, then a step that recalibrates brightness, color, etc. is used. Lastly, differences across the boundary (seam) are measured, and best candidates are chosen. This framework can also be used for de-identification.

Deep learning approaches include [31], which proposes a novel face swapping pipeline using a fully convolutional network for segmentations. Segmentation is preceeded by pose and expression estimation along with 3D shape fitting. The final step blends the source face into the target using the previously obtained segmentation. RS-GAN [29] uses two VAEs as the separator networks and one GAN that composes the separated face and the rest of the target image to generate the face-swapped output.

FSNet [28] uses a single VAE to disentangle the face apperance in the form of a latent vector. Then a GAN is used to combine the latent vector with the non-face part of the target image and synthesize the face-swapped result.

FSGAN [32] extends the framework with reenactment which is done by a novel recurrent neural network (RNN) that handles both pose and expression and can be also applied to videos. A GAN is used to inpaint the reenacted face to the target image and finally the blending is done by modifying Poisson blending with a novel loss. Recently, FSGANv2 [33] was proposed with multiple improvements in both architecture and experimental evaluation.

A novel attributes encoder and refinement network are proposed in FaceShifer [22]. The authors report a huge improvement in identity preservation and perceptual quality. RAFSwap [55] introduces a transformer [52] network that augments local identity-relevant features. They extract hierachical features using a pre-trained face parsing model and use an encoder to obtain the latent

representations of the images. This information is fused together with the transformer outputs and then fed to a StyleGAN2 [15] generator to produce the final image.

MegaFS [61] proposes the first megapixel level method for one shot face swapping. The authors create a Hierarchical Representation Face Encoder (HieRFE) that work in extended latent space to maintain more facial attributes. They also design a Face Transfer Module (FTM) that is used to transfer identity from the source to the target image. The final face synthesis is done using StyleGAN2 [15].

FSLSD [56] extends MegaFS with disentanglement of the identity and attributes in latent space and also extends the method to video face swapping.

Authors of DiffFace [16] propose the first face swapping model based on diffusion probabilistic models [46]. Specifically, they train identity-conditioned DDPM which is then used to denoise the target image while being conditioned on the identity from the source image.

2.3 Datasets

There are many public datasets consisting of face images. One dataset commonly used for training and verification of face anonymization models is the CelebA [24] dataset, or its variants like Celeb-DF [23]. The original CelebA consists of 202,599 images with a resolution of 178x218, and also includes annotations with 40 face attributes and 5 landmark locations.

CelebA-HQ is a high-quality version of CelebA consisting of 30,000 pictures with a 1024x1024 resolution. We used CelebA-HQ for some of our experiments in Chapter 5.

Labeled Faces in the Wild (LFW) [10] is another common dataset used in face anonymization research. LFW proposes a set of matching and non-matching pairs to evaluate face identification and verification methods. It consists of 13,233 images made from 5749 identities. We use LFW in some of our experiments available in Chapter 5.

Another dataset used for face anonymization is the People In Photo Albums (PIPA) [59]. WIDER FACE [57] has been used to benchmark the face detection performance of anonymization methods.

FFHQ [14] is a dataset similar to CelebA-HQ, with the same resolution of 1024x1024, and has been used for training of megapixel models. It consists of 70,000 images.

FaceForensics++ [40] contains 1000 real videos that have been used for testing of face anonymization in videos.

The authors of DeepPrivacy [12] propose the FDF dataset consisting of 1.47M human faces. The authors claim that larger and more diverse dataset enables their anonymization method to be more robust and realistic.

Chapter 3

Benchmark suite

Our proposed benchmark suite, AnonyBench, consists of 5 individual benchmarks that evaluate the quality of anonymization methods. In this chapter, we describe the implemented benchmarks and explain rationale behind them. We designed the benchmark suite based on existing research in the field and extended it with some more ideas of our own. The main goal is to obtain objective evaluation of any anonymization method with as much insight as possible. We implement the following 5 benchmarks that measure different aspects of the face anonymization methods:

1. Face detection - Measures whether the anonymized faces can still be detected. It serves as a proxy for facial distortion caused by anonymization.
2. Face re-identification - Computes distance between original and anonymized face in terms of identity. It measures how well the method actually anonymizes the original person.
3. Facial attributes - The goal is to measure how the anonymizer changes distribution of relevant face attributes like age, gender, and race. It is useful for recognizing various biases in anonymization methods.
4. GAN metrics - It computes commonly used metrics that proxy human image quality perception and are generally accepted for evaluation of generative models.
5. Anonymized images detector - Trains a linear SVM classifier using features extracted by a neural network pre-trained for face-related tasks. Measures how well the machine can distinguish between original and anonymized images.

Note that each benchmark can be run and evaluated independently, giving the user extended flexibility when using AnonyBench. We present the description and documentation of the actual software library in Appendix B. In the rest of the chapter, we thoroughly describe and discuss all the benchmarks.

3.1 Face detection

3.1.1 Motivation

Face detection is a task in computer vision that involves localizing human faces within images or videos. It is a necessary prerequisite for many applications, such as facial recognition or verification, facial attributes analysis, etc. We show an example of a successful face detection in Figure 3.1. Given the importance of face detection for other tasks, we would like that the anonymization process keeps the faces detectable.

In other words, this benchmark attempts to evaluate whether the anonymized faces can be detected as well as the original ones. This can be a proxy for measuring facial distortion caused by anonymization. This benchmark was proposed in [26] and extended in [19].

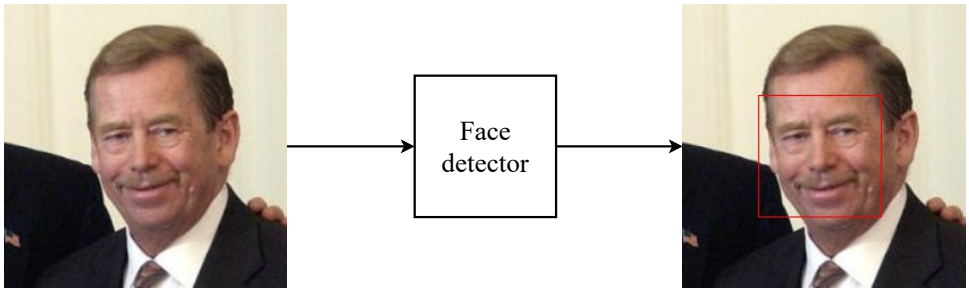


Figure 3.1: An example of a successful face detection.

3.1.2 Traditional anonymization methods

We also include several traditional anonymization methods as baselines. This is to show that traditional methods distort the faces and make them undetectable, thus reducing usefulness of the data for other downstream tasks.

The generation of the baseline datasets is automatically included in our suite. We show the whole range of evaluated methods in Figure 3.2.

We use 4 baseline methods:

- Black boxing - a black box over the face; the simplest baseline.
- Pixelization - the face area is downsampled 16 times and then upsampled back to original resolution resulting in a few large uniform boxes representing the downscaled pixels.
- Blur - a box blur with kernel 32x32 is applied to the face

- Full image blur - the same as blur but done to the whole image

One may notice that the baseline datasets have to be generated from the original data and a bounding box of the face has to be used. To make the benchmark as general as possible, we use the face detector that would be used for benchmarking the performance to extract the bounding boxes.

Given the fact that there may be original images where the face is not detectable, we omit those images from the benchmark, i.e. we only consider images where the face was detected in the original images.

3.1.3 Implementation

Since we only considered images in which the face was detected in the original picture, we use the **Fraction of Detected Faces (FoDF)** compared to the original dataset as the metric. We give the formula in Equation (3.1), where d_{orig} represents the number of images in the original dataset where the face was successfully detected, d_{anon} is the same metric for the anonymized dataset.

$$FoDF(d_{anon}, d_{orig}) = \frac{d_{anon}}{d_{orig}} \tag{3.1}$$

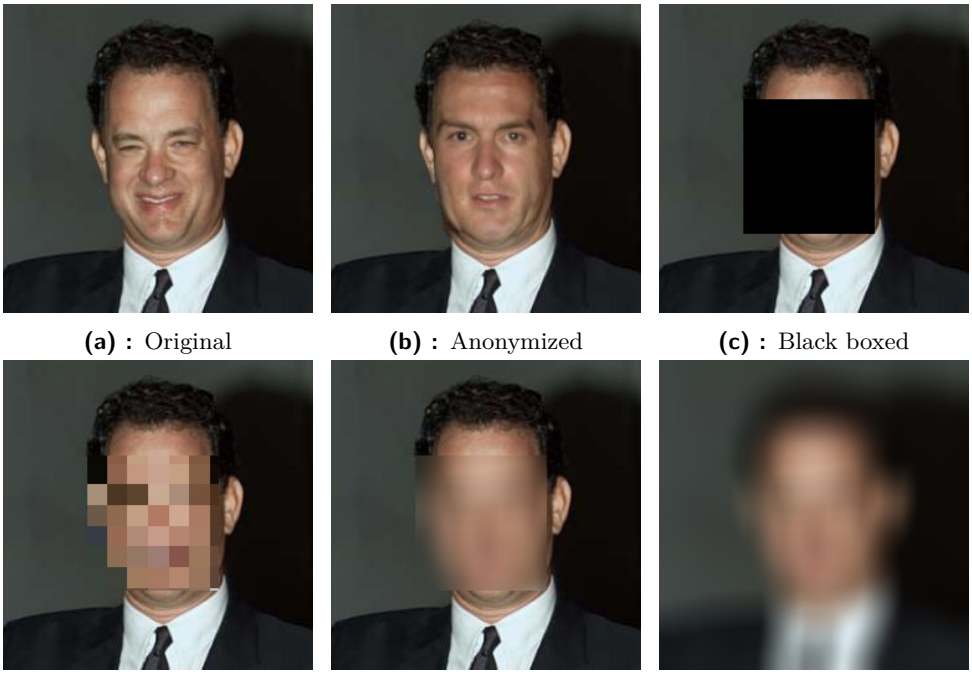
3.2 Face re-identification

3.2.1 Motivation

Face identification, also known as face recognition, is a technology that goes beyond face detection and focuses on recognizing and identifying specific individuals. Unlike face detection, which localizes faces in images or videos, face identification attempts to match a detected face to a specific identity.

The main goal of face anonymization is to change the face in a way that the resulting face's identity is different from the original one. This benchmark measures just that. We take each pair of original and anonymized images and compute distance between vector representations of their identity extracted by ArcFace [5].

We report the percentage of faces that were re-identified. We consider a face as re-identified if the distance between their identity vector representations is below the decision threshold. The threshold is usually computed using images of non-matching pairs of identities with a chosen false positive rate. The metric can be interpreted as error rate.



(a) : Original **(b) :** Anonymized **(c) :** Black boxed

(d) : Pixelized **(e) :** Blurred **(f) :** Full image blur

Figure 3.2: Showcase of anonymization methods evaluated in the face detection benchmark.

Note that trivial anonymizers, that provide heavily distorted faces, achieve high scoring according to the re-identification metric. That's the reason the evaluation has to take into account multiple metrics.

3.2.2 Implementation

We use the ArcFace [5] recognition model as a backbone of this benchmark. Specifically, ArcFace is used to extract embedding vectors from the faces. The embeddings are compared using cosine similarity. Since we want to measure the distance between the vectors, we use the **cosine distance**. We give the precise formula in Equation (3.2), given that $\mathbf{x}$ and $\mathbf{y}$ are the embedding vectors.

$$cosine\_distance(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{3.2}$$

The decision whether the anonymized face is considered re-identified is based on a distance threshold. We provide an option to compute this threshold using non-matching pairs of images and a false positive rate. If the non-matching pairs are not provided, a reasonable default threshold is used.

The decision threshold is set from the non-matching pairs using the following

strategy: We choose a false positive rate, e.g. 0.005. For each non-matching pair, we compute the distance of the vectors representing their identities. We order all the distances in ascending order. We set the threshold to make the lowest 0.5% of the distances fall below it and the rest be above it.

We show the complete pipeline of the face re-identification benchmark in Figure 3.3.

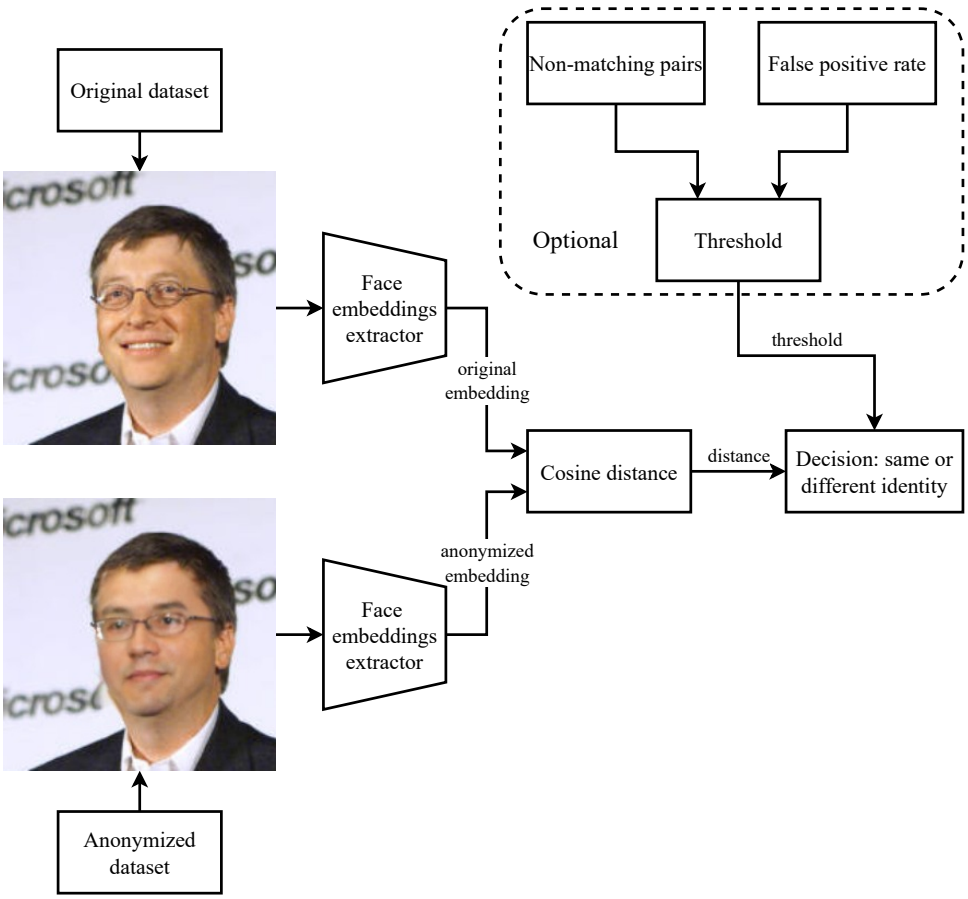


Figure 3.3: Complete pipeline of the face re-identification benchmark. Original and anonymized faces are converted to embeddings and cosine distance between them is computed. Final decision is made based upon a threshold that can be optionally computed from non-matching pairs of images with a given false positive rate.

3.3 Facial attributes

3.3.1 Motivation

In Chapter 1, we define the task of face anonymization as the generation of a new face that should preserve the original facial attributes but have a

different identity. In this benchmark, we measure whether facial attributes were preserved. Specifically, we compare age, gender and race of original and anonymized faces.

These attributes help us to identify potential biases in the methods. Our experiments show that some methods show strong biases with respect to facial attributes, further information and discussion are presented in Chapter 5. We note that the benchmark for facial attributes has not been explored in previous work, thus it is a novel contribution. We show an example of the facial attribute analysis process in Figure 3.4.

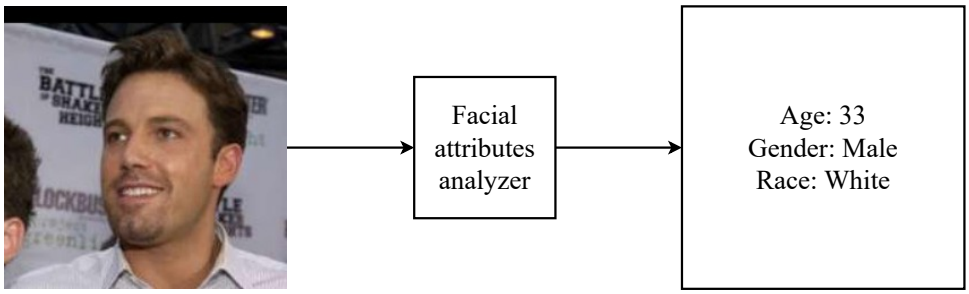


Figure 3.4: An example of facial attributes analysis.

3.3.2 Implementation

To objectively compare the methods, we propose a single-valued metric for each of the measured attributes. We use age, gender, and race models from the DeepFace [45] library.

Age. We compute the absolute age difference for each pair of images. The resulting aggregated metric is the mean of absolute age differences. This metric captures the average change in age between original and anonymized faces, giving insight into age preservation.

We give the the precise formula for **Mean absolute age difference (MAAD)** in Equation (3.3), where x_i^{orig} denotes the estimated age of the original face and x_i^{anon} represents the same for the anonymized face, n is the number of pairs in the dataset. We note that some authors refer to this as the Mean Absolute Error (MAE).

$$MAAD(x_1^{orig}, x_1^{anon}, \ldots, x_n^{orig}, x_n^{anon}) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{orig} - x_i^{anon}\right| \tag{3.3}$$

Gender and race. Since both gender and race attain finite values, we consider them as classification tasks. The metric of choice for both attributes is the preservation. For each class, we compute probability that the anonymized

class is the same as the original one. In terms of general classification metrics, this corresponds to **recall**.

Note that recall is computed for each class individually. To create a single metric, we take an arithmetic mean of the recalls for each class (gender or race). The arithmetic mean is unweighted; all classes have the same weight to promote gender and race fairness. We give the formula for **preservation** in Equation (3.4), where C represents a set of classes (either genders or races) and *anon*, *orig* represent the estimate of the given attribute from the anonymized, original face.

$$preservation(C) = \frac{1}{|C|} \sum_{c \in C} P(anon = c \mid orig = c) \tag{3.4}$$

In the above equation, we use the probabilistic interpretation of recall, in other words, what is the probability that an anonymized face would have attribute c given that the original face had attribute c. More specifically, the conditional probability is given in Equation (3.5), where I_c is a set of indexes of faces, which have attribute c in the original image.

$$P(anon = c \mid orig = c) = \frac{1}{|I_c|} \sum_{i \in I_c} \delta(c_i^{orig} = c_i^{anon}) \tag{3.5}$$

In addition to the single-valued metrics proposed, we also show the histogram of differences for the age metric. For gender and race, we also report the actual confusion matrices.

3.4 GAN metrics

3.4.1 Motivation

Since the inception of generative models, researchers have been trying to create a metric that can be used to judge the quality of the generated samples. Correlation with human judgement is a vital requirement for such metric.

Over the past years, many metrics have been proposed. Even though these metrics only serve as a proxy to human judgement, they are still quite useful for relative comparisons of different generative models, which is the reason we include this benchmark in the suite.

This benchmark uses several generally accepted metrics to evaluate the quality of anonymized images. The metrics used have been shown to correlate with human image quality assessments. Specifically, we used 3 metrics:

- Fréchet Inception Distance (FID) [9]
- Learned Perceptual Image Patch Similarity (LPIPS) [60]
- Structural Similarity Index Measure (SSIM) [53]

In the rest of this section, we describe each metric in more detail, unless explicitly stated, we follow the notation from the original articles. Note that all of these metrics are commonly used in the literature on generative models and are not exclusively related to face anonymization.

Note that FID computes the distance between two distributions, which is the reported value. On the other hand, LPIPS and SSIM compute similarity between a pair of images. The reported value for them is the mean of pairwise similarities.

3.4.2 Fréchet Inception Distance (FID)

FID was originally proposed in [9] as an improvement to Inception score [41].

Inception-V3 [51] network is used to summarize each image into a feature vector x, specifically the coding layer, which is the last pooling layer before the classifier part of the network. FID only considers the first two moments: mean and covariance. The assumption is made that the distribution is multidimensional Gaussian. The distance of the Gaussians is measured by the Fréchet distance, i.e., the Wasserstein-2 distance.

$(\mathbf{m}, \mathbf{C})$, $(\mathbf{m}_w, \mathbf{C}_w)$ are estimated using features extracted through the Inception-V3 network. $(\mathbf{m}, \mathbf{C})$ are computed from the generated data, while $(\mathbf{m}_w, \mathbf{C}_w)$ are computed from the original data. The FID is then given by [6]:

$$d^2((\mathbf{m}, \mathbf{C}), (\mathbf{m}_w, \mathbf{C}_w) = \|\mathbf{m} - \mathbf{m}_w\|_2^2 + Tr\left(\mathbf{C} + \mathbf{C}_w - 2(\mathbf{C}\mathbf{C}_w)^{1/2}\right) \quad (3.6)$$

The authors of [9] then show that FID is consistent with human judgement. We note that FID is the most common generative model quality metric used for the evaluation of face anonymization methods.

3.4.3 Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [60] attempts to create a metric for the evaluation of the perceptual similarity between two images. Authors show that using deep features outperforms all previous metrics by a large margin. The actual distance is calculated using AlexNet [20] as a feature extractor.

The two evaluated images (or image patches) x, x_0 are converted to features, the features are unit normalized in the channel dimension. Activations of

each layer are then scaled by a vector. Note that LPIPS computes the metric by taking ℓ_2 distance and averaging spatially and summing channel-wise.

The formula adopted from [60] is given in Equation (3.7), where $\hat{y}^l$, $\hat{y}^l_0$ represent features extracted from layer l, H_t, W_t represent height and width of the layer l, and w_l represents the scaling weights for the layer l.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})\|_2^2 \tag{3.7}$$

3.4.4 Structural Similarity Index Measure (SSIM)

SSIM is a traditional method for assessment of perceptual image quality. It is based on the concept of degradation of structural information. The task is separated into three comparisons: luminance, contrast, and structure.

The luminance is computed as the mean intensity given by Equation (3.8). The comparison between two images is then a function $l(\mathbf{x}, \mathbf{y})$ of μ_x and μ_y.

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.8}$$

Next, SSIM removes the mean intensity from the image $\mathbf{x} - \mu_x$.

The signal contrast is estimated by the standard deviation, given by Equation (3.9).

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{\frac{1}{2}} \tag{3.9}$$

The contrast comparison $c(\mathbf{x}, \mathbf{y})$ is then the comparison between σ_x and σ_y. The structure comparison $s(\mathbf{x}, \mathbf{y})$ is done on normalized images $(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$.

The overall similarity measure is given:

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y}))$$

where

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

and C_1, C_2, and C_3 represent small constants to avoid numerical instability. We combine the equations into SSIM as follows:

$$SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y}) \cdot c(\mathbf{x}, \mathbf{y}) \cdot s(\mathbf{x}, \mathbf{y})$$

Finally, SSIM is given by Equation (3.10), where we used $C_3 = C_2/2$.

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.10}$$

We note that in case of SSIM, higher value means higher similarity, thus *higher is better*.

3.5 Anonymized images detector

3.5.1 Motivation

The last proposed benchmark trains a classifier that acts as a detector of anonymized pictures. The main motivation is to evaluate how well a machine can distinguish between original and anonymized pictures.

In case of perfect anonymization and perfectly balanced data (i.e. 50% real vs. 50% anonymized), we would expect the accuracy of the detector to be around 50%, which means that the original and anonymized images are indistinguishable and the detector would randomly guess the label.

On the other hand, if the anonymization is extremely poor, we would expect the accuracy to reach 100%, i.e. detector could perfectly distinguish between original and anonymized images.

3.5.2 FaceNet

FaceNet [43] is a generic embedding model for face-related tasks. It uses an InceptionResNet-V1 model [50] pre-trained on CASIA-Webface [58]. FaceNet is an efficient model that extracts good face representations that are later used to train the linear SVM classifier.

3.5.3 Implementation

Since original and anonymized dataset have the same size, we are dealing with a balanced binary classification problem. Our pipeline is as follows:

1. We extract the features using FaceNet, which is a pre-trained deep network for face-related tasks.
2. We use the extracted features to train a linear SVM classifier.
3. We evaluate the classifier on a test set, specifically, we report the classification accuracy, computed as the ratio of correctly predicted images.

Note that SVM training is done using k-fold cross-validation with 5 folds to get a better statistical estimate of the accuracy. Specifically, we use 4 folds for training and 1 fold for testing. We fix the the SVM regularization constant C to 1. We show a diagram of the detector pipeline in Figure 3.5.

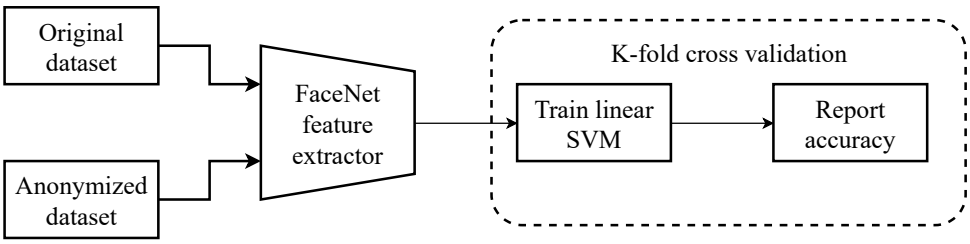


Figure 3.5: The pipeline of the anonymized images detector benchmark. Features are extracted from datasets using FaceNet [43]. Then k-fold cross validation is employed to train a linear SVM using the extracted features. Finally, accuracy is reported as the metric.

Chapter 4

Proposed face swapping to anonymization conversion

While evaluating existing anonymization methods, we discovered that inpainting-based methods exhibit strong biases in terms of facial attributes, for example, an anonymized face has a much lower apparent age or different apparent race.

In this chapter, we propose a novel approach to face anonymization by employing a face-swapping model, facial attributes analyzer, and an annotated database of faces.

We show that our method is competitive to the existing inpainting-based anonymization methods and strongly improves upon biases in facial attributes. This is because the proposed method explicitly controls the attribute distribution of the generated images, which is possible due to the usage of face swapping model combined with a facial attributes analyzer.

We carried out experiments using the CelebA-HQ [24] dataset. We use MegaFS [61] as the backbone face-swapping model (see Section 4.2 for more information about MegaFS). We present the results of the experiment in Section 5.2. Note that we refer to our method as **AnonySwap**.

4.1 Implementation

To employ face-swapping model as an anonymizer, we need a database of faces that can be swapped into the target image. Since MegaFS natively works with CelebA-HQ, we used CelebA-HQ as our face database. Moreover, CelebA-HQ has annotations for identities.

Identites are useful for the re-identification benchmark, we can easily compute a decision threshold from the non-matching pairs. Moreover, the identity can be used as an additional criterion when using CelebA-HQ as a

database to anonymize images from CelebA-HQ, i.e. we know the identity for the input face and will only choose candidates for face swapping that do not represent the same identity.

For each face in the database, we run the DeepFace [45] facial attribute analysis model. Specifically, we estimate age, gender, and race. We store these annotations along with the faces.

We make sure that all possible combinations of genders and races are available in the database. When searching for candidates to face swap, first we select the images matching gender and race. Then we select candidates based on age. We start with an absolute age difference of 0 and continue to iteratively increase the difference by 1 until at least one candidate is found. Finally, we randomly select one candidate from the set of possible candidates.

Our approach also allows to implant a specific identity. The only change would be a slight modification in the candidate selection method.

Now, we are ready to anonymize new images, the process is as follows:

1. Age, gender, and race of the input face is estimated.
2. We search for faces with matching facial attributes in the database.
3. One face out of the matching set is chosen randomly.
4. We swap the chosen face onto the input image.

We show an example diagram of the described process in Figure 4.1.

4.2 MegaFS

Megapixel level method for one shot Face Swapping (MegaFS) [61] is the face swapping model we used to test our anonymization framework. It is made up of three independent components:

1. **Hierarchical Representation Face Encoder (HieRFE).** HieRFE is used to project the two face images into the latent space. The image is encoded into three parts: constant StyleGAN2 input denoted as C, four latent codes representing low-level information denoted as L^{low}. Finally, other latent codes representing high-level semantic information are gathered as L^{high}.
2. **Face Transfer Module (FTM).** FTM combines L^{high} of both faces into a new one representing the result of face swapping. Let L_t^{high}, L_s^{high} represent the high-level latent codes for the source and target images,

then FTM produces L_{s2t} that represents high-level latent codes for the resulting image.

3. **Pretrained StyleGAN2 generator.** Lastly, pretrained StyleGAN2 [15] is used to generate the swapped face image from latent codes. Specifically C_t and L_t^{low} represent previously mentioned constant StyleGAN2 input and low-level latent codes for the target image. They're used along with L_{s2t} to synthesize the swapped face.

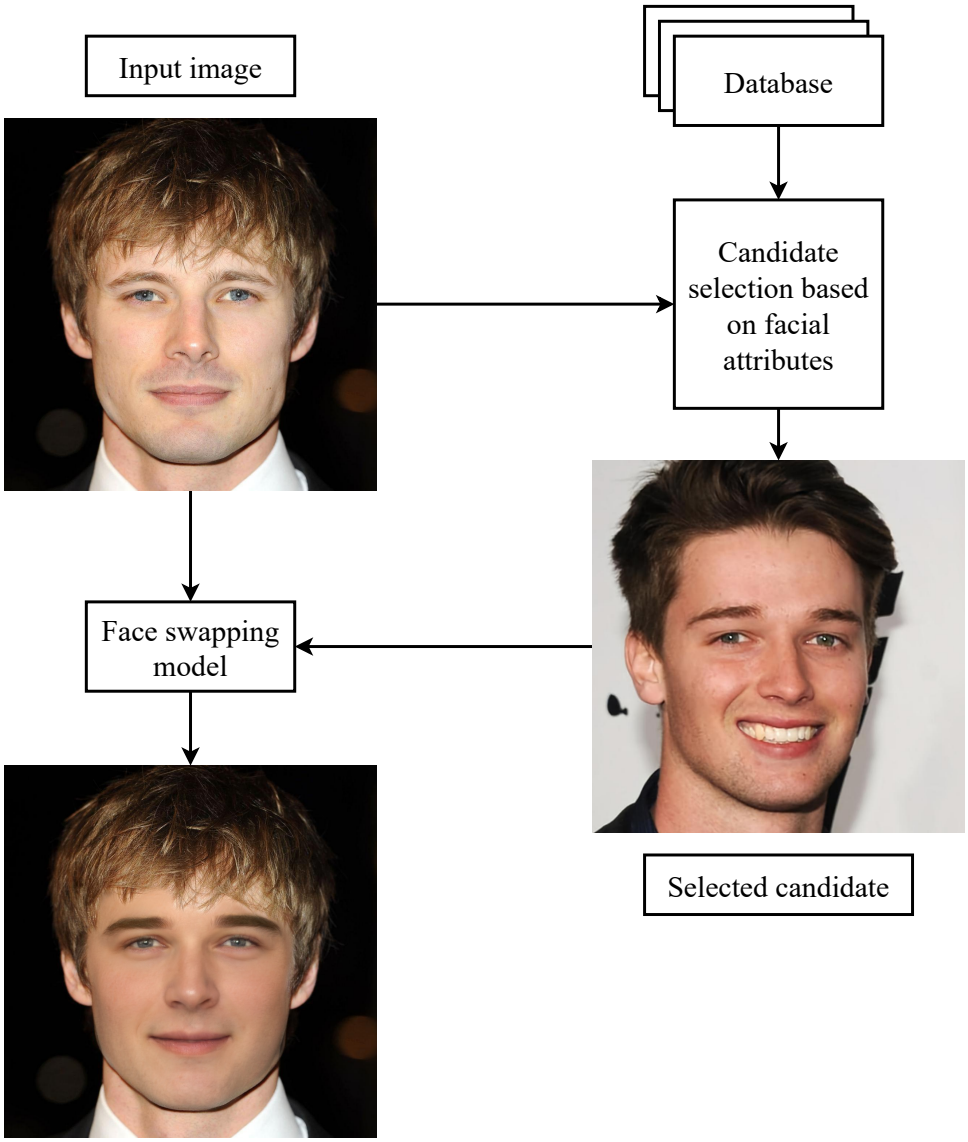


Figure 4.1: An example pipeline of anonymization via face swapping. Facial attributes from the input image are used to find a suitable candidate in the database. Then a face swap is performed between the input image and the selected candidate.

Chapter 5

Experiments

In this chapter, we evaluate several face anonymization methods using CelebA-HQ and LFW datasets and we perform a human study that evaluates the photorealism of several anonymization methods. We start with an overview of the competing methods in Section 5.1.

Next, we evaluate three anonymization methods on the CelebA-HQ dataset. We present the results in Section 5.2. The purpose of this evaluation is to show competitiveness our proposed framework from Chapter 4. We use CelebA-HQ dataset because the face-swap model MegaFS [61] is programmed to work with CelebAMask-HQ [21] dataset. We use MegaFS as a backbone for our framework AnonySwap.

We conducted our main evaluation by applying five anonymization methods to the LFW dataset. In Section 5.3, we discuss the results of the evaluation along with a final ranking table computed from the average performance of all the benchmark criteria.

Lastly, we examine the results of the human study in Section 5.4. The study was done using the LFW dataset. We used samples from the five methods evaluated in Section 5.3.

5.1 Competing methods

In this Section, we give an overview of the evaluated methods. Note that the general description of MegaFS [61] is available in Section 4.2.

5.1.1 CIAGAN

CIAGAN [26] uses a conditional GAN to generate a new face. The generator is implemented as an encoder-decoder U-Net [39]. It is conditioned on landmarks

of the original face along with the background from the original image, the face itself is blacked out.

Finally, an identity is chosen from a pre-trained set. The identity is encoded as a one-hot vector and fed into the generator bottleneck.

Together with a standard GAN discriminator, the authors introduce an identity discriminator that guides the generator to synthesize faces according to the chosen identity.

5.1.2 DeepPrivacy

DeepPrivacy [12] proposes a conditional GAN conditioned on pose information. Specifically, 7 keypoints describe the pose of the face; including the left and right shoulder.

The generator uses a modified U-Net [39] architecture. The pose information is concatenated with output after each upsampling layer of the decoder.

The discriminator is modified to include the background information (image without the face) as conditional input, i.e. the input image has six channels instead of three. Pose information is concatenated with the input of each downsampling layer of the discriminator.

The network is trained using a custom Flickr Diverse Faces (FDF) dataset that consists of 1.5M faces *in the wild*.

5.1.3 DeepPrivacy2

DeepPrivacy2 [11] improves the original DeepPrivacy by introducing a new GAN model with higher resolution (256x256). Moreover, the new model does not use pose information.

The decoder of the generator follows the design of StyleGAN2 [15]. The authors also introduce the FDF256 dataset, which is a subset of the FDF used in DeepPrivacy made from higher-quality images.

5.1.4 AnonySwap + FSGAN

FSGANv2 [33] is a face swapping method. Therefore, we make use of our proposed framework from 4. We will refer to the method as *FSGAN* rather than *FSGANv2* to avoid clutter.

FSGANv2 uses a recurrent neural network (RNN) for reenactment. It takes the source face and changes both pose and expression of the face to

match the target face. A U-Net-based [39] model is used for the segmentation of the target image.

An inpainting GAN is used to combine the reenactment and segmentation outputs. Finally, a blending step is applied to merge the inpainted face with the target image.

5.1.5 LDFA

LDFA [18] is a diffusion-based model that uses inpainting weights of Stable Diffusion [38]. The inpainting weights are general, i.e. they are not specifically trained for face synthesis. The Euler Ancestral sampler is used with 50 inference steps. No prompt is used with CFG-scale of 1.

5.2 Evaluation using CelebA-HQ

The main motivation behind using face-swap models for face anonymization is the improved preservation of facial attributes. We evaluate this criterion and some others using our proposed benchmarking suite. Note that all benchmarks are thoroughly described in Chapter 3.

We use the CelebA-HQ [24] dataset for our experiments (MegaFS internally employs the CelebAMask-HQ [21] dataset). We show some example pictures from the dataset in Figure 5.1.

The competing methods are the state-of-the-art inpainting-based methods DeepPrivacy [12] and DeepPrivacy2 [11]. We chose these two methods because they achieve the best performance in our human study (Section 5.4).

5.2.1 Facial attributes

First, we evaluate performance on the facial attributes benchmark, as this benchmark has been the main motivation for the new anonymization method.

The results obtained by AnonySwap + MegaFS show greatly increased performance in terms of race preservation, substantial improvement in gender preservation, and lower mean absolute age difference when compared to the competing inpainting-based methods.

We show the results in Table 5.1. One may notice that race preservation of around 57% is still far from 100%. We attribute this to two factors. First, the actual model for race estimation is not perfect and correctly classifying race is a complex problem.

Figure 5.1: Example images from the CelebA-HQ dataset.

Second, MegaFS is still far from perfect in terms of visual quality of produced images. Therefore, it can, e.g. produce a result with non-matching illumination of the face that may further confuse the race estimation model. Same can be said for the estimation of other attributes.

Method	MAAD(↓)	Gender preservation(↑)	Race preservation(↑)
DeepPrivacy	5.33 ± 4.58	81.46%	33.24%
DeepPrivacy2	4.88 ± 4.42	87.55%	31.35%
AnonySwap+MegaFS	**3.71 ± 3.83**	**91.54%**	**57.27%**

Table 5.1: Facial attributes benchmark on CelebA-HQ: results show that our proposed method outperforms existing inpainting-based anonymization methods.

5.2.2 Face re-identification

The performance on the face re-identification benchmark is less impressive, as shown in Table 5.2. While the re-identification percentage is low, inpainting anonymization methods perform better. We attribute this to the fact that face-swap models work with the identity of the original face. Therefore, some attributes of the original face may be preserved in the final result.

Method	Re-identified faces(↓)
DeepPrivacy	3.15%
DeepPrivacy2	**3.08%**
AnonySwap + MegaFS	8.97%

Table 5.2: Face re-identification benchmark on CelebA-HQ: results show that inpainting-based methods still take the edge off our proposed method. This is due to the fact that face swap models work with the identity of original face. Therefore, some identity-related attributes may be preserved in the final result. We used ArcFace [5] with false positive rate of 0.5%. We used pairs of non-matching identities from CelebA-HQ to estimate the decision threshold using the given false positive rate.

5.2.3 GAN metrics

Table 5.3 shows that FID, LPIPS, and SSIM of our method is the lowest, which means the best, among the compared methods. We attribute this to the high quality of the pre-trained StyleGAN2 [15] generator.

Method	FID(↓)	LPIPS(↓)	SSIM(↑)
DeepPrivacy	28.9693	0.2943	0.8329
DeepPrivacy2	14.9329	0.2277	0.8436
AnonySwap + MegaFS	**14.391**	**0.1508**	**0.9084**

Table 5.3: GAN metrics benchmark on CelebA-HQ: results show that our method outperforms existing anonymization methods.

5.2.4 Anonymized images detector

Lastly, the anonymized images detector benchmark results show that it's harder to learn a good detector for our proposed method than it is for others, as shown in Table 5.4. On the other hand, we can see that the accuracy is quite high, which shows that anonymized images are still easily detectable by a machine.

Method	Accuracy(↓)
DeepPrivacy	99.50%
DeepPrivacy2	94.55%
AnonySwap + MegaFS	**92.81%**

Table 5.4: Anonymized images detector benchmark on CelebA-HQ: results show that our method achieves lowest accuracy, therefore it's harder to separate original and anonymized images than for other methods.

5.3 Evaluation using LFW

We believe that our proposed suite of benchmarks can be used to efficiently evaluate new or existing face anonymization methods. In this section, we compare five face anonymization methods on the Labeled Faces in the Wild (LFW) [10] dataset. The evaluated methods are described in Section 5.1. We also refer the reader to Appendix A for all the plots concerning this evaluation.

We chose the LFW dataset because it contains identity annotations. Threshold for face re-identification is computed using non-matching pairs of images that can be obtained using the identity annotations.

Another advantage of the LFW dataset is the lower image quality (250x250) that makes the images more realistic and fuzzy, which is better for human-based evaluation (see Section 5.4) since it makes shortcomings of the methods less visible. We show some example images in Figure 5.2. We also show sample images with their anonymized counterparts in Figure 5.3.

Figure 5.2: Example images from the LFW dataset.

5.3.1 Face detection benchmark

We show the results of the face detection benchmark in Table 5.5. We can see that the only underperforming method is LDFA; we attribute this to the usage of a general inpainting model that can generate something else than a face.

(a) : Original **(b) :** CIAGAN **(c) :** DeepPrivacy **(d) :** DeepPrivacy2 **(e) :** FSGAN **(f) :** LDFA

Figure 5.3: Sample original images and their anonymized versions.

When we compare the results with the traditional methods in Table 5.6, we can notice how pooly traditional methods perform.

5.3.2 Face re-identification benchmark

The face re-identification benchmark results (see Table 5.7) show that AnonySwap does not change the identity well enough to be used for anonymization. A histogram of the distances for AnonySwap is shown in Figure A.4.

On the other hand, CIAGAN achieves the best result. We observed that faces produced by CIAGAN have low degree of photorealism. We hypothesize that low photorealism puts the generated faces further away from the realistic ones in the embedding space. We show a histogram of the distances in Figure A.1.

Method	Detected faces(↑)	Rank(↓)
CIAGAN	99.22%	3
DeepPrivacy	99.04%	4
DeepPrivacy2	99.43%	2
AnonySwap + FSGAN	**99.73%**	**1**
LDFA	93.07%	5

Table 5.5: Face detection benchmark on LFW: Percentage of detected faces after anonymizing the original data.

Method	Detected faces(↑)
Boxed	6.28%
Blurred	12.31%
Pixelized	6.57%
Full blur	7.71%

Table 5.6: Face detection benchmark on LFW: Percentage of detected faces after anonymizing the original data using baseline methods.

5.3.3 Facial attributes benchmark

Table 5.8 shows the results of the facial attributes benchmark. AnonySwap achieves the best performance. This is not surprising, since AnonySwap is the only face-swap model in the comparison and we have shown that face swapping does a better job at preserving facial attributes than inpainting-based methods.

We show the MAAD histogram in Figure A.9. We can see that AnonySwap will generally make a person look younger rather than older.

CIAGAN is the worst performing method. Given the architecture that inpaints a set of pretrained identities, we would expect poor results in this regard.

We note that plots of relative confusion matrices are available in Appendix A.

5.3.4 GAN metrics benchmark

We show the computed GAN metrics in Table 5.9. We can see that DeepPrivacy2 wins this benchmark. There are large discrepancies in FID score. We believe that the reason may be the necessary image resizing for CIAGAN, AnonySwap, and LDFA.

For other metrics, we can see that only CIAGAN is really poor, while other methods perform reasonably well.

Method	Re-identified faces(↓)	Rank(↓)
CIAGAN	**2.08%**	**1**
DeepPrivacy	8.35%	4
DeepPrivacy2	6.70%	3
AnonySwap + FSGAN	41.46%	5
LDFA	4.29%	2

Table 5.7: Face re-identification benchmark on LFW: Percentages of re-identified faces. We used ArcFace [5] with false positive rate of 0.5%. We used pairs of non-matching identities from CelebA-HQ to estimate the decision threshold using the given false positive rate.

Method	MAAD(↓)	Gender preservation(↑)	Race preservation(↑)	Ranks(↓)
CIAGAN	7.01 ± 5.92	66.81%	21.97%	5, 5, 5
DeepPrivacy	5.69 ± 4.97	89.88%	35.73%	2, 2, 2
DeepPrivacy2	6.04 ± 5.23	83.02%	27.22%	3, 3, 4
AnonySwap+FSGAN	**5.13 ± 4.90**	**94.49%**	**60.04%**	**1, 1, 1**
LDFA	6.66 ± 5.72	82.90%	32.87%	4, 4, 3

Table 5.8: Facial attributes benchmark on LFW: Results show that our proposed method outperforms other methods.

Method	FID(↓)	LPIPS(↓)	SSIM(↑)	Ranks(↓)
CIAGAN	15.1677	0.3545	0.4137	5, 5, 5
DeepPrivacy	2.4355	0.0724	0.8902	2, 3, 3
DeepPrivacy2	**1.7853**	**0.0543**	0.8914	**1**, **1**, 2
AnonySwap + FSGAN	11.6778	0.0592	**0.8986**	4, 2, **1**
LDFA	11.6339	0.0928	0.8374	3, 4, 4

Table 5.9: GAN metrics benchmark on LFW: DeepPrivacy2 obtains best results.

5.3.5 Anonymized images detector benchmark

The results of the anonymized images detector benchmark are presented in Table 5.10. We notice that the rankings of this benchmark are very similar to those of the GAN metrics benchmark. DeepPrivacy2 achieves best performance, which means it is hardest to separate original and anonymized images.

Method	Accuracy(↓)	Rank(↓)
CIAGAN	98.24%	5
DeepPrivacy	88.71%	3
DeepPrivacy2	**86.36%**	**1**
AnonySwap + FSGAN	88.64%	2
LDFA	93.77%	4

Table 5.10: Anonymized images detector benchmark on LFW: detector accuracy when trying to distinguish real and anonymized images.

5.3.6 Final rankings

Finally, we take an average of the achieved ranks for all the criteria and assign a final rank to each method based on the average. We also include the date published for fairness. See Table 5.11.

It is evident that both DeepPrivacy methods and AnonySwap are competitive, in contrast to CIAGAN and LDFA, which perform poorly. When comparing the results to the human study from Section 5.4, it becomes apparent that AnonySwap does not perform well.

We hypothesize that this is due to the fact that face swapping methods still struggle with the visual quality of generated samples, while inpainting methods perform much better in that regard. However, AnonySwap emerges as the clear winner in terms of preserving facial attributes.

Method	Published	Average rank(↓)	Final rank(↓)
CIAGAN	November 2020	4.33	5
DeepPrivacy	October 2019	2.78	3
DeepPrivacy2	January 2023	2.22	2
AnonySwap + FSGAN	February 2022	**2.00**	**1**
LDFA	February 2023	3.67	4

Table 5.11: Final rankings table. Average rank is computed as an unweighted mean of ranks attained for all measured criteria. We can notice that both DeepPrivacy methods and AnonySwap are competetive, while CIAGAN and LDFA perform worse.

5.4 Human annotation study

Finally, we conducted a human study. Each annotator labeled 150 images from the LFW dataset and had to classify the image as real or generated. As for the 150 images; 75 images are real, 75 are generated using one of the evaluated anonymization methods (15 images for each method).

We collected 3150 labels from 21 annotators. We show the results in Table 5.12. DeepPrivacy and DeepPrivacy2 achieve substantially better results than other methods, i.e. they are much harder to detect by humans.

We also note that **12.95%** of the original images were labeled as generated by the annotators.

Method	Recall(↓)	Rank(↓)
CIAGAN	98.73%	5
DeepPrivacy	**42.86%**	**1**
DeepPrivacy2	50.48%	2
AnonySwap + FSGAN	92.06%	4
LDFA	91.43%	3

Table 5.12: Human annotator recall when trying to distinguish between real and generated images.

Chapter 6

Conclusions

In this thesis, we investigated the field of face anonymization, focusing on deep generative models.

First, we developed a suite for benchmarking of face anonymization methods. One of the main outputs of the thesis is a CLI tool that can be used to objectively evaluate and compare different methods for face anonymization. The benchmark suite consists of five individual benchmarks. The CLI library offers a simple yet configurable interface, which is convenient to use and user-friendly.

We also created a general framework that can be used to convert the task of face anonymization to the task of face swapping. Poor preservation of facial attributes was the motivation behind this framework. We showed that our proposed approach substantially improves the performance in that regard.

However, face swapping methods still face challenges in achieving photorealistic quality in the resulting images. In this case, inpainting-based methods perform better.

Lastly, we used our suite to evaluate several existing methods, along with a human study. The results of the human study indicate that annotators have around 50% recall, i.e. they are guessing randomly, when labeling images anonymized using DeepPrivacy or DeepPrivacy2. Nevertheless, these methods are strongly biased in terms of facial attributes.

Our evaluation reveals significant potential for future improvements. We hope that our work can provide insight into the strengths and weaknesses of individual methods and help in future research.

In summary, in this work, we have developed a benchmarking suite for face anonymization methods, which is runnable as a CLI tool along with another CLI tool for visualization. We have proposed a framework that converts the task of face anonymization to the task of face swapping to improve preservation of facial attributes. Finally, we evaluated several existing methods along with

a human study.

Future work

As the scope of this work is large, there are many possibilities for future work. One possible direction is to implement new benchmarks extending the current set. Some improvements can also be made to the developed CLI tools.

We showed that best methods in terms of visual perception have strong biases with respect to facial attributes. On the other hand, non-biased methods are not as well received in terms of visual quality. A new model that could perform well in both regards would be an interesting research direction.

Training a diffusion-based model specifically for face anonymization could be another interesting area of research. We note that such an endeavor would take a lot of computational resources.

Another possibility for future work is to extend our proposed AnonySwap method by trying out new models, datasets, or candidate selection methods.

Glossary

CLI Command-line interface. 39, 40, 63, 64

FID Fréchet Inception Distance. 18, 31

GAN Generative Adversarial Network. 5–7, 27–29, 34, 63

GDPR General Data Protection Regulation. 1

LFW Labeled Faces in the Wild. ix, x, 8, 27, 32, 34–36, 51–61

LPIPS Learned Perceptual Image Patch Similarity. 18, 19, 31

MAAD Mean absolute age difference. 16, 30, 34, 35

SSIM Structural Similarity Index Measure. 6, 18–20, 31

VAE Variational Auto-Encoder. 6, 7

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27, August 2008.

[3] Durkhyun Cho, Jin Han Lee, and Il Hong Suh. CLEANIR: Controllable Attribute-Preserving Natural Identity Remover. *Applied Sciences*, 10(3):1120, February 2020.

[4] Nicola Dall'Asen, Yiming Wang, Hao Tang, Luca Zanella, and Elisa Ricci. Graph-based generative face anonymisation with pose preservation. In *Image Analysis and Processing – ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*, page 503–515, Berlin, Heidelberg, 2022. Springer-Verlag.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[6] D.C Dowson and B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[8] R. Gross, L. Sweeney, F. De La Torre, and S. Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 161–161, New York, NY, USA, 2006. IEEE.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Gary Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Tech. rep.*, October 2008.

[11] H. Hukkelas and F. Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338, Los Alamitos, CA, USA, jan 2023. IEEE Computer Society.

[12] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[13] Itseez. Open source computer vision library. `https://github.com/itseez/opencv`, 2015.

[14] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228, dec 2021.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[16] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. DiffFace: Diffusion-based Face Swapping with Facial Guidance, December 2022.

[17] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[18] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. LDFA: Latent Diffusion Face Anonymization for Self-driving Applications, February 2023.

[19] S. R. Klomp, M. Van Rijn, R. J. Wijnhoven, C. M. Snoek, and P. N. De With. Safe fakes: Evaluating face anonymizers for face detectors. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[21] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, Seattle, WA, USA, June 2020. IEEE.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation, October 2021.

[26] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455, June 2020.

[27] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. Face Deidentification with Generative Deep Neural Networks. *IET Signal Processing*, 11(9):1046–1054, December 2017.

[28] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*, 2018.

[29] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. *RSGAN: Face Swapping and Editing Using Face and Hair Representation in Latent Spaces*. Association for Computing Machinery, New York, NY, USA, 2018.

[30] E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, February 2005.

[31] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 98–105, Los Alamitos, CA, USA, may 2018. IEEE Computer Society.

[32] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019.

[33] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGANv2: Improved subject agnostic face swapping and reenactment. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2022.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[36] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11205, pages 639–655. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[37] Slobodan Ribaric and Nikola Pavesic. An overview of face de-identification in still images and videos. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Ljubljana, May 2015. IEEE.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[40] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[41] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[42] Branko Samarzija and Slobodan Ribaric. An approach to the de-identification of faces in different poses. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1246–1251, Opatija, Croatia, May 2014. IEEE.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[44] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[45] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[47] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5050–5059, 2017.

[48] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 570–586, Cham, 2018. Springer International Publishing.

[49] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, 2022.

[50] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

[51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[53] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[54] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-Protective-GAN for Face De-identification. *Journal of Computer Science and Technology*, pages 47–60, 2019.

[55] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7622–7631, 2022.

[56] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022.

[57] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning Face Representation from Scratch, November 2014.

[59] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir D. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. *CoRR*, abs/1501.05703, 2015.

[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[61] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun. One shot face swapping on megapixels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4832–4842, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

Appendix A

Experiment plots

We show all plots related to experiments on the Labeled Faces in the Wild (LFW) dataset. We choose to separate them from the main experiments in Section 5.3 to avoid clutter. Specifically we show plots for:

- Histogram of distances for face re-identification benchmark in Figures A.1, A.2, A.3, A.4, A.5.
- Histogram of age differences for facial attributes benchmark in Figures A.6, A.7, A.8, A.9, A.10.
- Relative gender confusion matrix for facial attributes benchmark in Figures A.11, A.12, A.13, A.14, A.15.
- Relative race confusion matrix for facial attributes benchmark in Figures A.16, A.17, A.18, A.19, A.20.

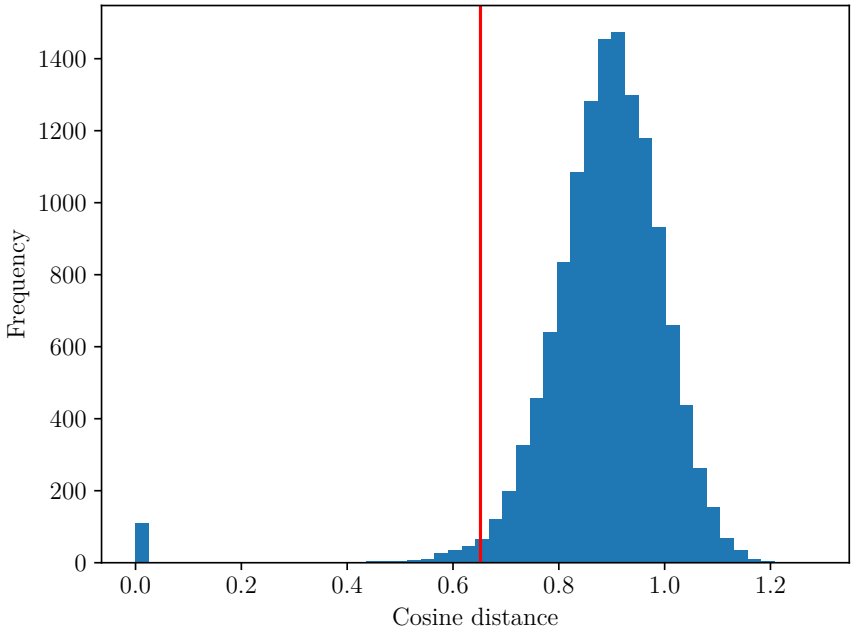


Figure A.1: Histogram of face re-identification distances for CIAGAN on LFW dataset. Red line represents the decision threshold for re-identification.

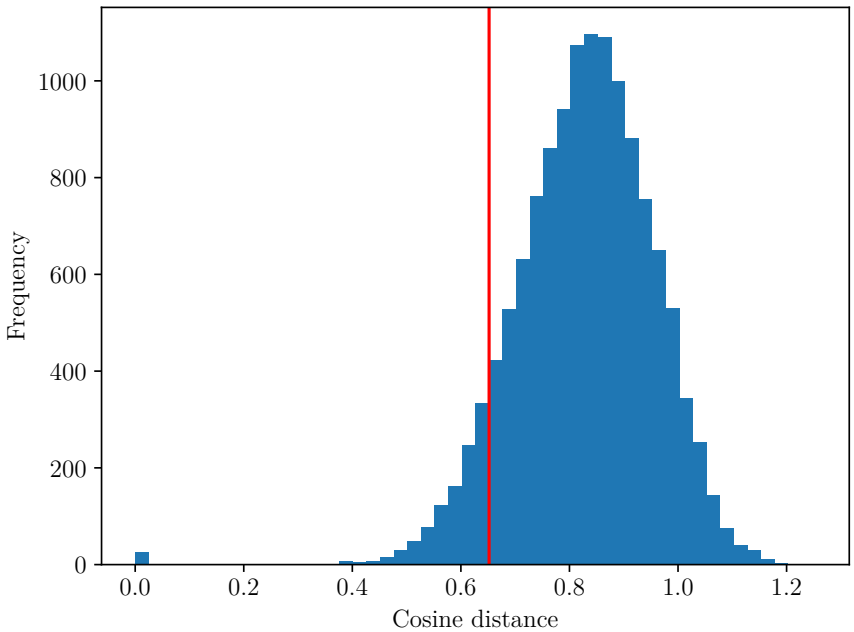


Figure A.2: Histogram of face re-identification distances for DeepPrivacy on LFW dataset. Red line represents the decision threshold for re-identification.

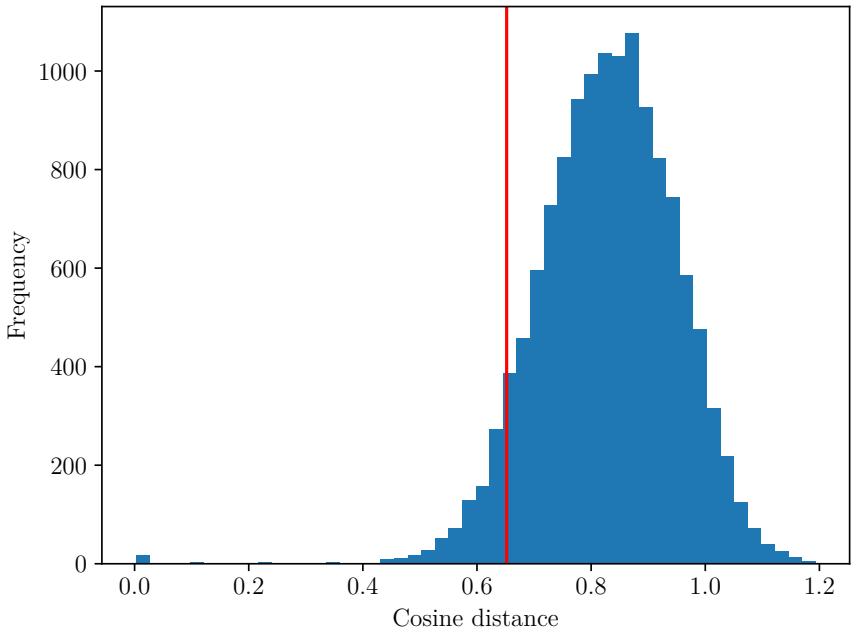


Figure A.3: Histogram of face re-identification distances for DeepPrivacy2 on LFW dataset. Red line represents the decision threshold for re-identification.

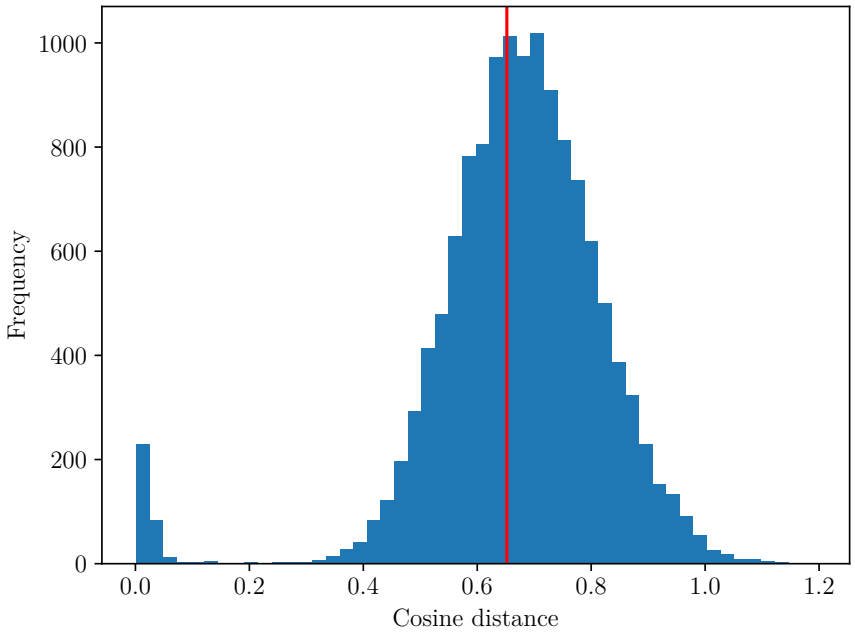


Figure A.4: Histogram of face re-identification distances for AnonySwap + FSGAN on LFW dataset. Red line represents the decision threshold for re-identification.

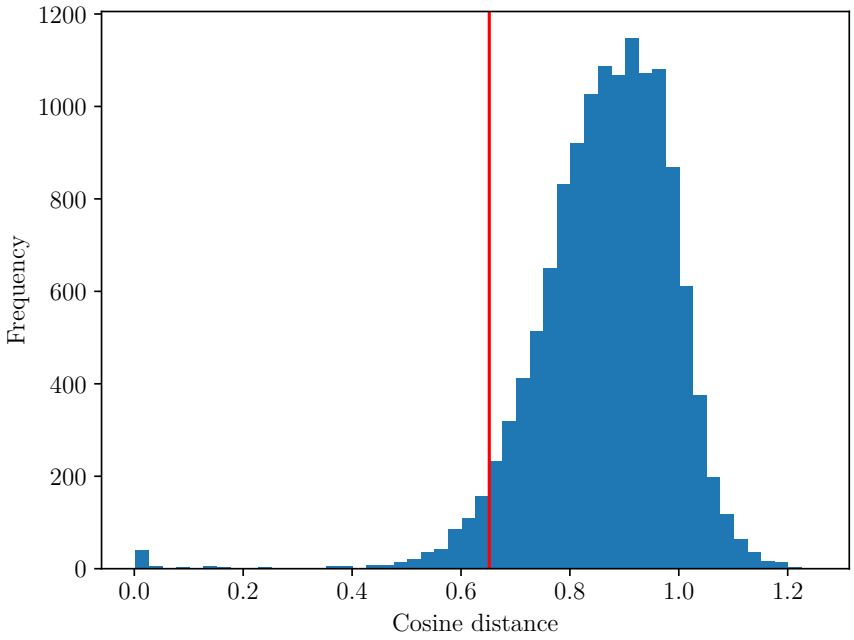


Figure A.5: Histogram of face re-identification distances for LDFA on LFW dataset. Red line represents the decision threshold for re-identification.

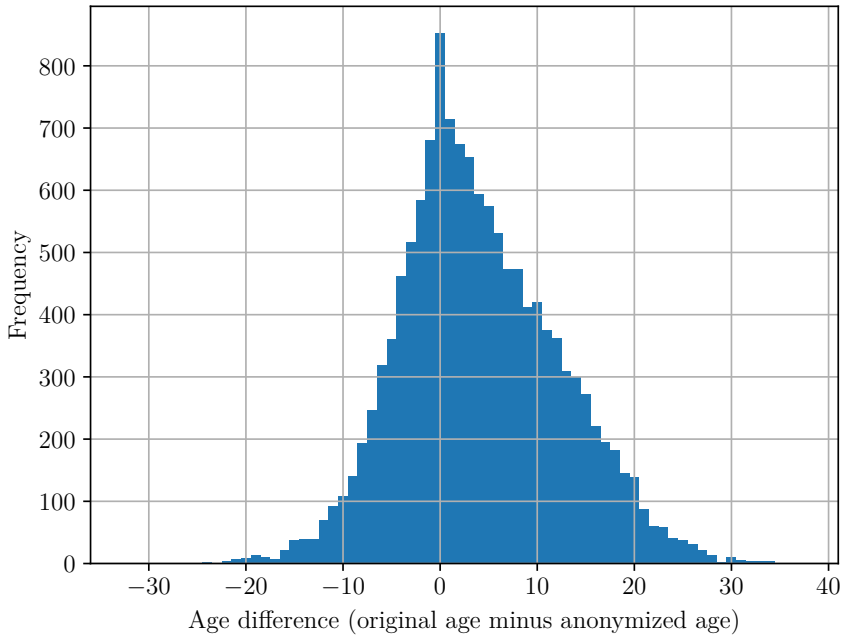


Figure A.6: Histogram of age differences for CIAGAN on LFW dataset.

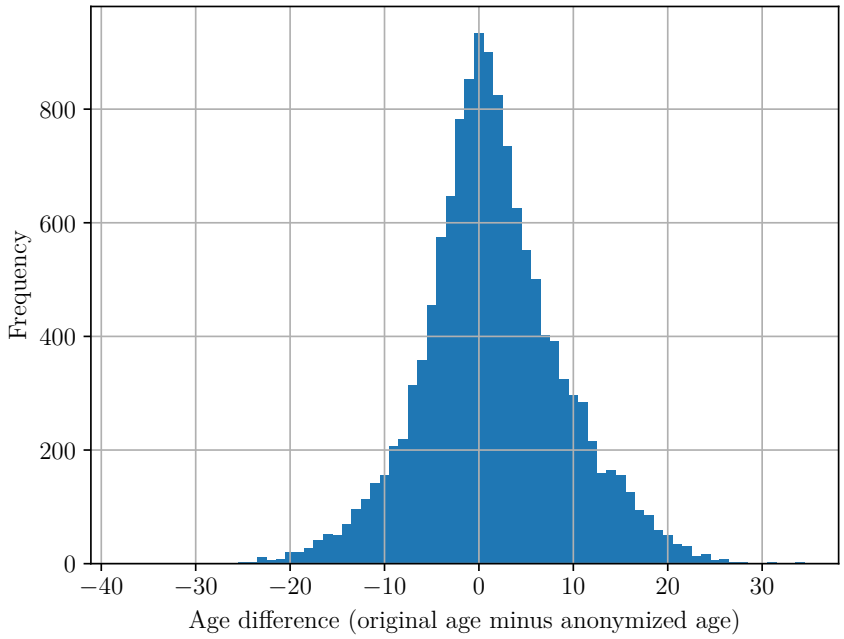


Figure A.7: Histogram of age differences for DeepPrivacy on LFW dataset.

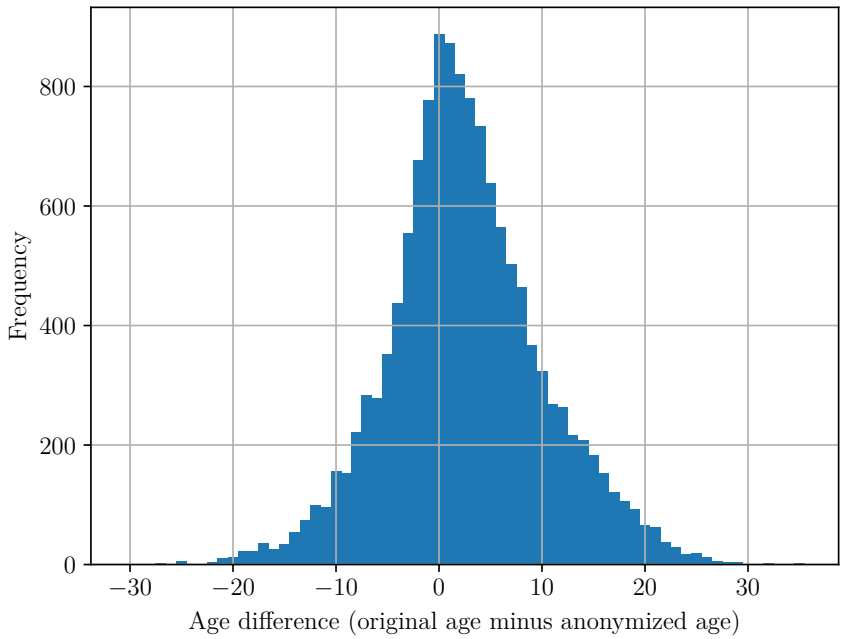


Figure A.8: Histogram of age differences for DeepPrivacy2 on LFW dataset.

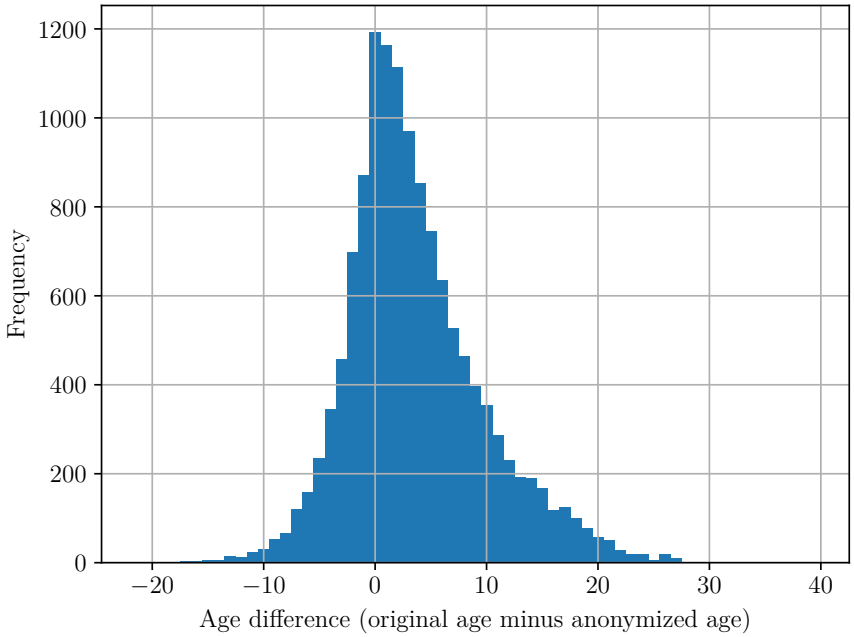


Figure A.9: Histogram of age differences for AnonySwap + FSGAN on LFW dataset.

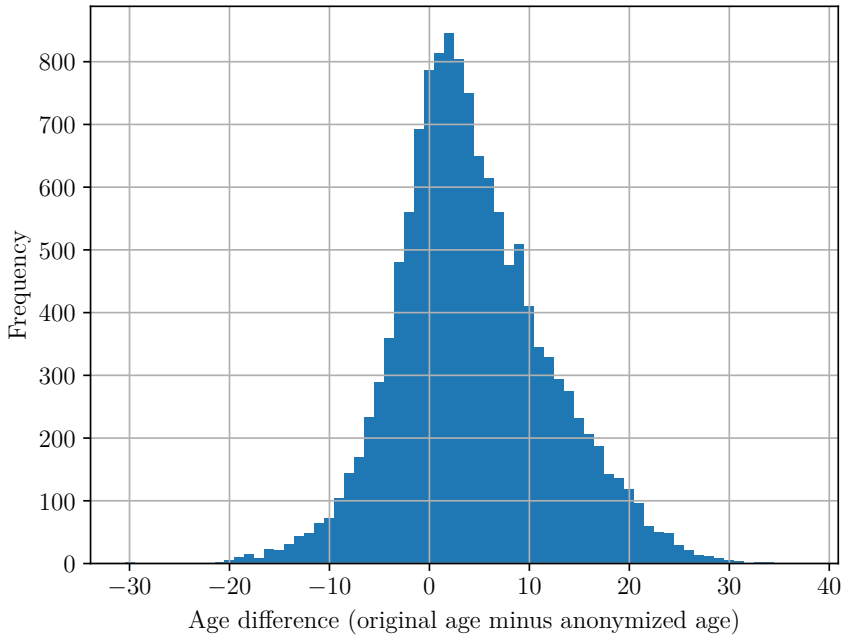


Figure A.10: Histogram of age differences for LDFA on LFW dataset.

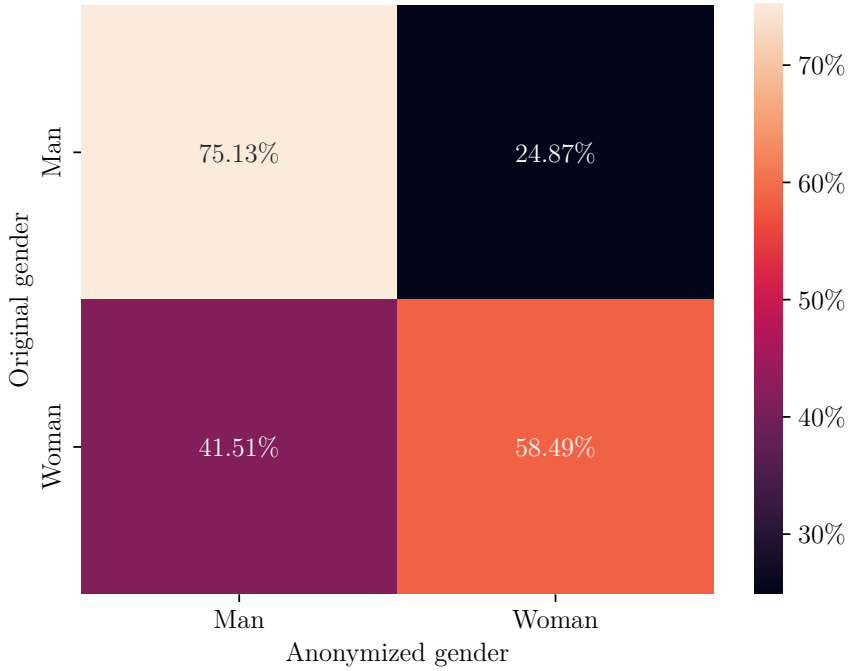


Figure A.11: Relative gender confusion matrix for CIAGAN on LFW dataset.

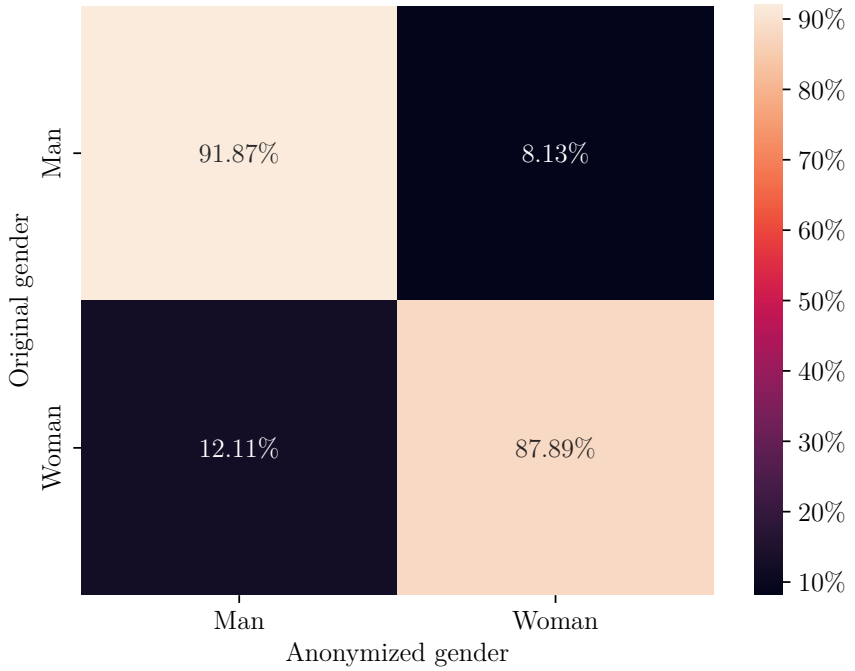


Figure A.12: Relative gender confusion matrix for DeepPrivacy on LFW dataset.

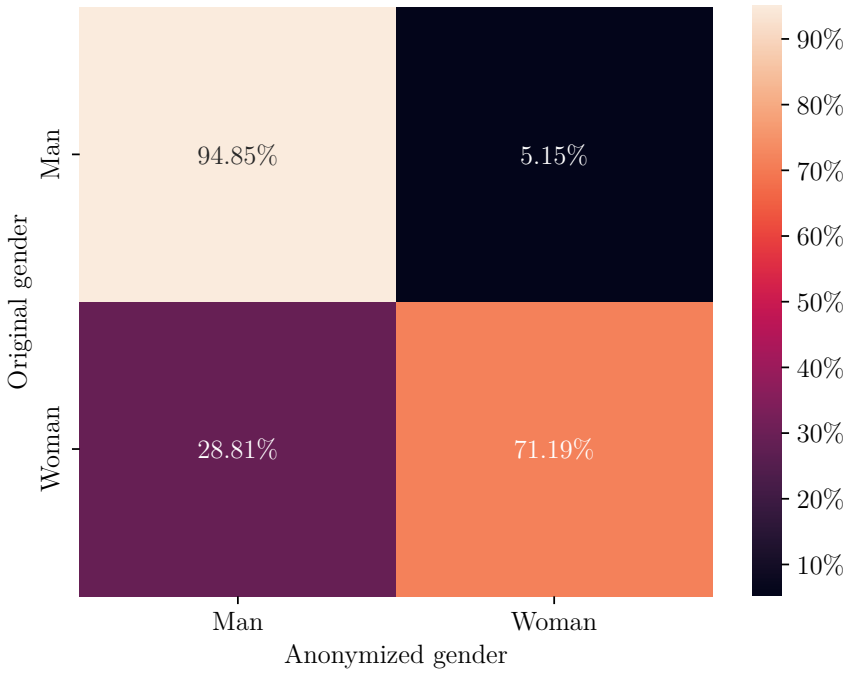


Figure A.13: Relative gender confusion matrix for DeepPrivacy2 on LFW dataset.

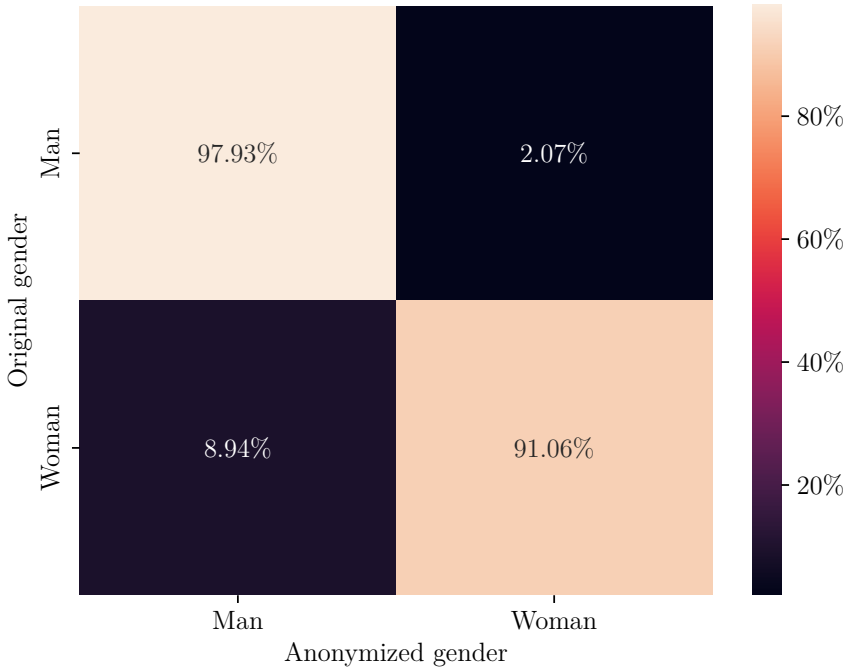


Figure A.14: Relative gender confusion matrix for AnonySwap + FSGAN on LFW dataset.

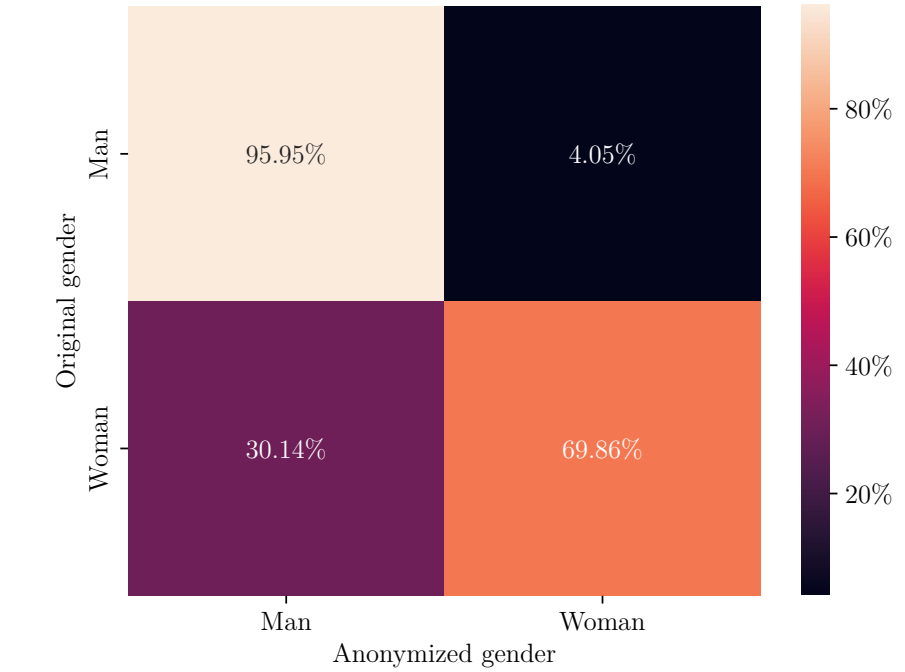


Figure A.15: Relative gender confusion matrix for LDFA on LFW dataset.

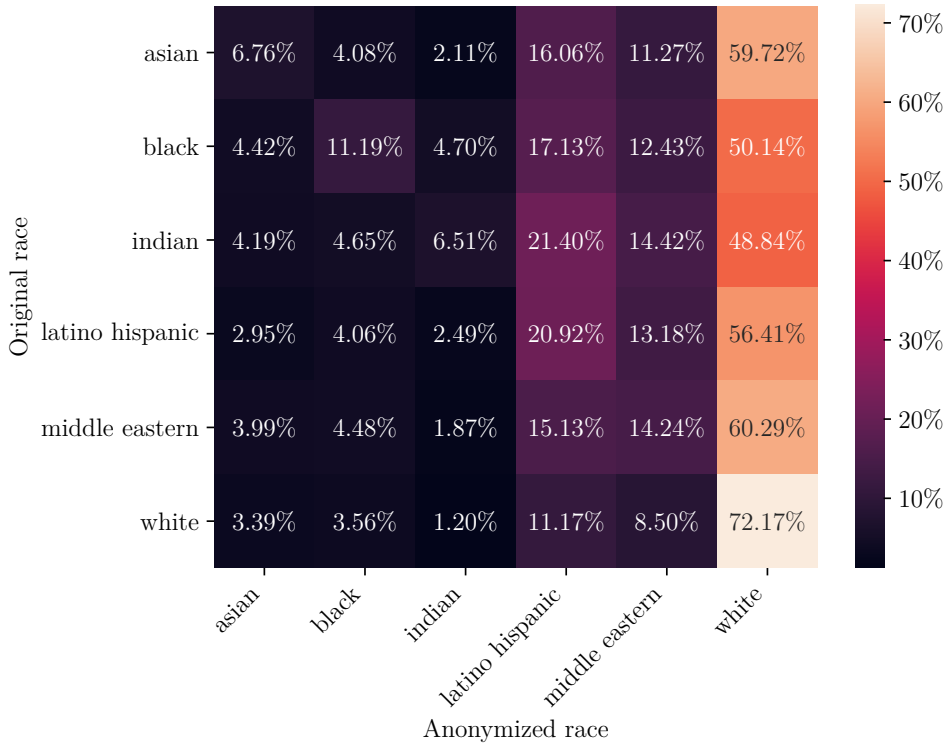


Figure A.16: Relative race confusion matrix for CIAGAN on LFW dataset.

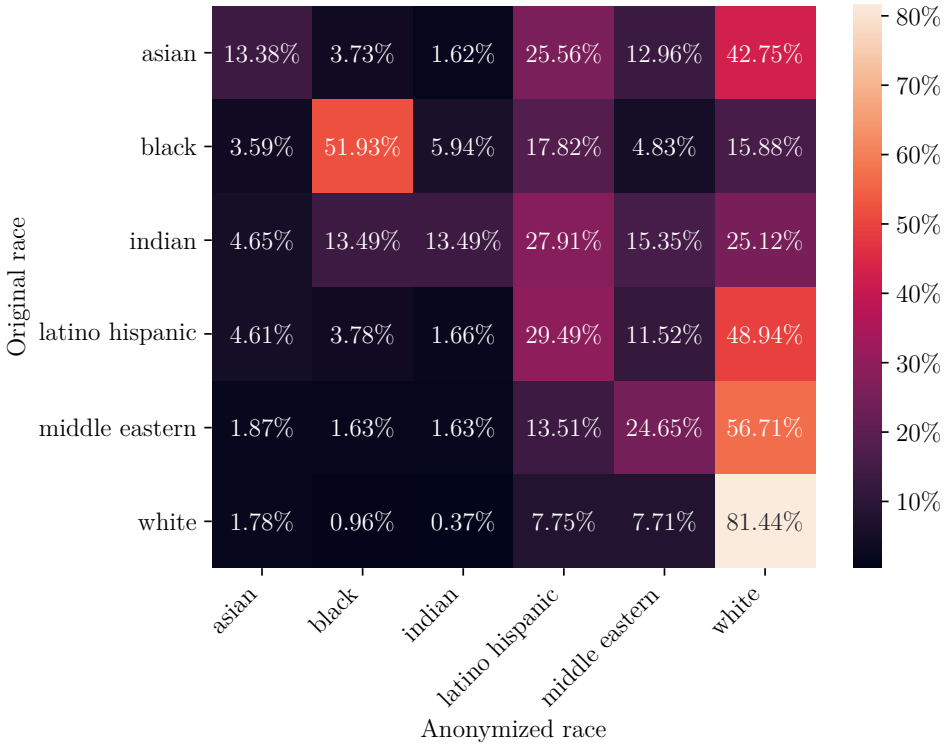


Figure A.17: Relative race confusion matrix for DeepPrivacy on LFW dataset.

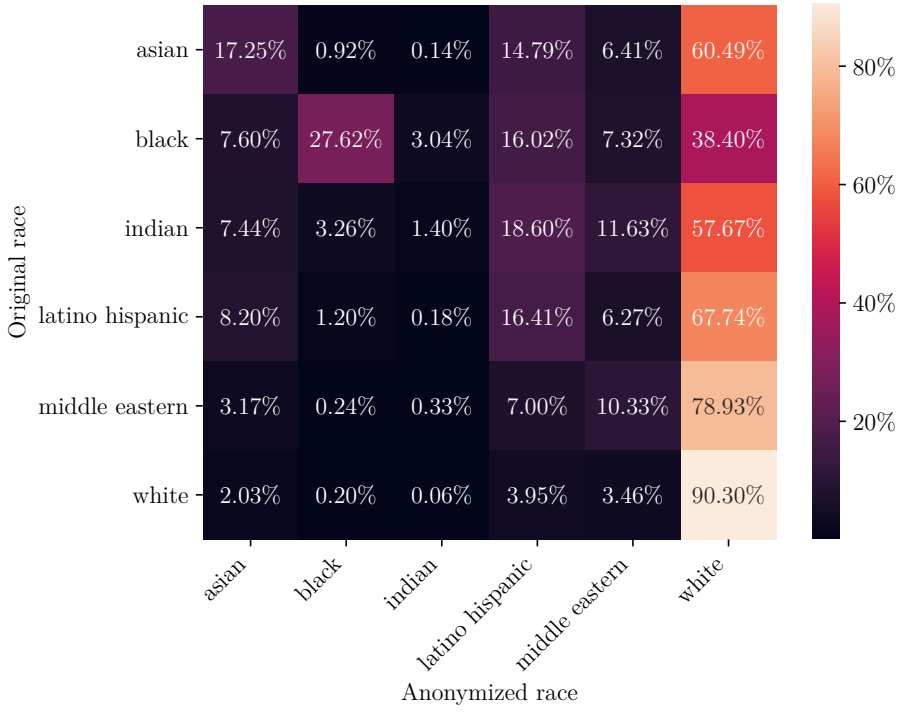


Figure A.18: Relative race confusion matrix for DeepPrivacy2 on LFW dataset.

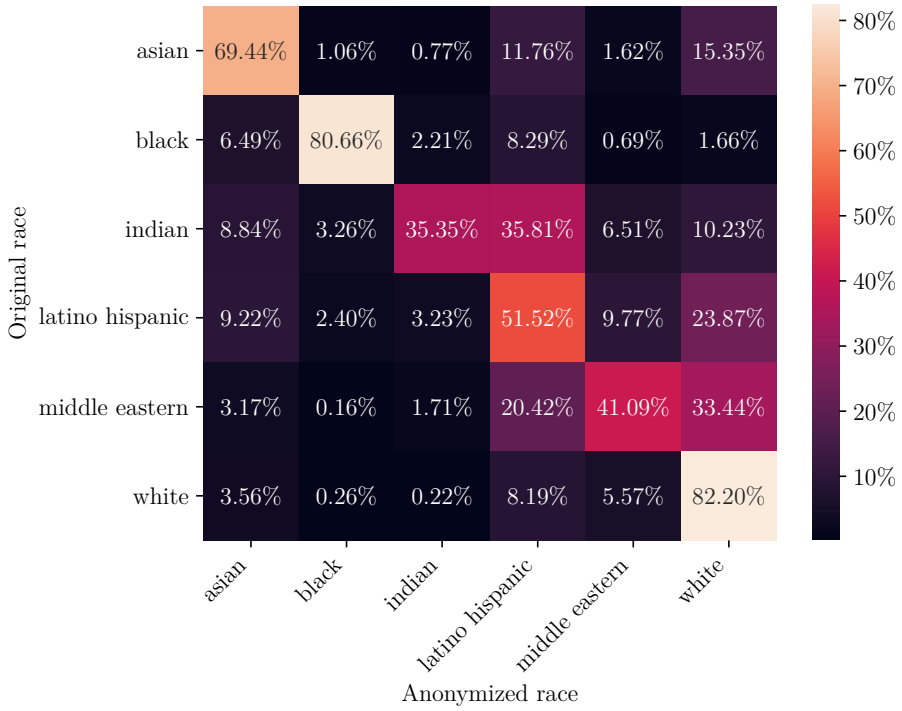


Figure A.19: Relative race confusion matrix for AnonySwap + FSGAN on LFW dataset.

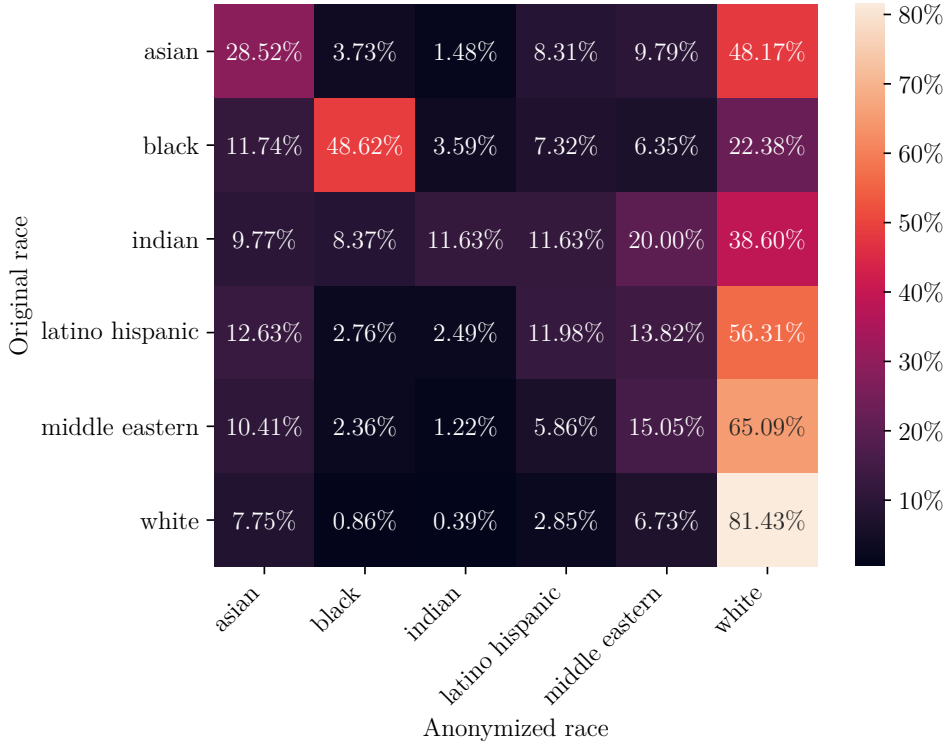


Figure A.20: Relative race confusion matrix for LDFA on LFW dataset.

Appendix B

Software library description

In this chapter, we describe the proposed benchmarking software library. Our suite consists of 2 command-line utilities. The first Command-line interface (CLI) runs the actual benchmark suite, while the second CLI runs the visualization based on the results of the first CLI. The library is available at https://github.com/jirimoravcik/AnonyBench.

We developed the whole suite in Python, specifically version 3.8. Both CLI utilities are made using the `argparse` module from the Python standard library. We use dlib [17] for face detection.

For image manipulation, we use OpenCV [13]. For deep learning models for face identification and facial attributes, we use DeepFace [44] [45] that uses TensorFlow [1] under the hood. We use PyTorch-based libraries [34] for GAN metrics: torchmetrics and clean-fid.

For the anonymized images detector benchmark, we used facenet-pytorch package for the FaceNet model and the sklearn package for the linear SVM classifier.

B.1 Benchmarks CLI

We provide a CLI to run the benchmarks. The CLI has two required arguments:

- `original_dataset_dir` is a path to the directory with the original dataset.
- `anonymized_dataset_dir` is a path to the directory with the anonymized dataset.

While these 2 parameters are enough to run the benchmark suite, we provide more optional parameters to provide more robustness and user-friendliness:

- `-h`, `--help` prints a help message with all possible arguments to the cli, help messages for each argument and default values for optional arguments.
- `-g`, `--use_gpu` switches computation of TensorFlow and PyTorch to GPU if possible.
- `-v`, `--verbose` makes the logger print more information to the console.
- `-o`, `--output_dir` sets the output folder for benchmark results, i.e. where will be the `.csv` outputs stored.
- `-b`, `--benchmarks` selects which benchmarks to run. If not specified, the whole suite is ran.
- `--fp_rate` sets the false positive rate for face re-identification benchmark
- `--batch_size` defines the batch size for deep learning models. Decreasing the batch size may help with memory issues.
- `--non_matching_pairs_filepath` gives a file of non-matching pairs from which the face re-identification threshold is computed, if omitted, a reasonable default is used.

B.2 Visualization CLI

We also provide a CLI for the visualization of benchmark results. We provide two formats: `.txt` and `.html`. The text format is better for development and quick results. However, the HTML format provides a richer experience and several plots that are not available in the text version.

Although CLI does not have any required arguments, there are several optional arguments:

- `-h`, `--help` prints a help message with all possible arguments to the cli, help messages for each argument and default values.
- `-s`, `--source_dir` modifies the source directory from which the results are taken. By default, this is set to the default output directory of the benchmark CLI.
- `-v`, `--verbose` makes the logger print more information to the console.
- `-o`, `--output` changes the file name of the output file with visualization.
- `-f`, `--folder` changes the folder where output file with visualization will be stored.

- `-f`, `--format` defines the output format, this can be `html`, `latex`, or `txt`.
- `-n`, `--name_contains` defines that only files including the given string will be present in the visualization