FLAG LAB

Dissertation

Fine-Tuning Face Anti-Spoofing Models: Exploring the Transformative Impact of Image-Level Biometric Privacy-Enhancing Techniques

Fernando Avalos

February 2, 2024

This thesis is submitted in partial fulfillment of the requirements for a degree of Bachelors in Systems and Computing Engineering (ISIS).

Thesis Committee:

Prof. Rubén MANRIQUE (Advisor) Universidad de los Andes, Colombia

Fine-Tuning Face Anti-Spoofing Models: Exploring the Transformative Impact of Image-Level Biometric Privacy-Enhancing Techniques

Systems and Computing Engineering Department

FLAG lab

Faculty of Engineering

Universidad de los Andes

Bogotá, Colombia

This manuscript has been set with the help of TEXSHOP and PDFLATEX (with BIBTEX support). Bugs have been tracked down in the text and squashed thanks to *Bugs in Writing* by **?**] and *Elements of Style* by **?**].

"Study hard what interests you the most in the most undisciplined, irreverent and original manner possible." - Dr. Richard Feynman

Abstract

Face Anti-Spoofing (FAS) systems are entrusted with the task of determining whether the content of a facial image is genuine or fake. The performance and overall quality of those systems depend on the data they are fed during their development stage, meaning that high-resolution and diverse facial datasets are commonly used. Even though FAS systems are particularly useful in some situations, they do raise concerns in regards to the privacy of the individuals whose images are used to train them. The straightforward solution is to apply filters to the facial images, at the expense of the systems' performance, since recognising faces becomes increasingly harder. This is where Biometric Privacy-Enhancing Techniques (B-PETs) come into play, which help to alleviate the adversarial tension between biometric utility and privacy gains. This work concerned itself with assessing the impact of 3 distinct image-level B-PETs in the performance of 3 architecturally distinct FAS systems and found that image-level B-PETs are not fit for finding a valuable trade-off, suggesting that more sophisticated techniques are needed.

Acknowledgements

I would like to express my most sincere gratitude to my advisor Rubén Manrique, who gave me insightful advice to carry out my thesis and that remained patient despite some of my missteps.

I would also like to thank all of those in my immediate social circle, who provided me tremendous amount of support throughout all the brainstorming, execution and writing processes this workpiece entailed. Particularly, I thank my parents, Paula and Manuel and my two best friends, Olga and Juan Sebastián.

Contents

List of Figures

Chapter 1

Introduction

In the recent years, seemingly dissimilar albeit related (in a contrived manner) technologies have arised. In one hand, we have the rise of highly specialized Artificial Intelligences, some of then capable of generating images that wouldn't be recognized as artificial by humans. On the other, we have the steady rise IoT, promising an even tighter integration between mass consumption technology and the real world. What could these two have in similar? They are, in some way, a threat to Face Anti-Spoofing (FAS) systems, in the sense that both technologies can be used to trick the aforementioned systems; the former technology provides a relatively straightforward way to generate high-fidelity face images of people of interest to gain unauthorized access over a system while the second might be used as a vector for extensive behavioural and contextual information mining regarding face images.

Given such threats, the need to enrich and strengthen FAS systems is stronger than ever. It has been shown that data augmentation from real world face datasets can lead to improved performance of FAS systems [11]. It could be contended that the baseline problems is that *real* world data is a requirement for fine-tuning FAS systems, conflicting with another relevant concept: *privacy*.

Privacy as a concept is not entirely free from controversy, but given that it has been conceptualised since the Classics [10] were alive and it is mentioned in the Universal Human Rights Declaration, one could assume that protecting privacy in all the stages involved in the development of FAS systems is a non-negotiable requirement.

That fact alone motivated the need for the so-called Biometric Privacy-Enhancing Techniques

(B-PETs), which ought to find a balance between concealing information from potentially malicious third-party actors and retaining some degree of utility for developing robust and competent FAS systems. As such, natural questions that could be asked are: What are these techniques? How are they applied? How do they impact the performance of existing FAS systems? Can the balance between privacy-enhancement and biometric utility be quantified? If so, how? These paper aims to answer these and similar questions in the context of three already existing FAS models.

Chapter 2

Literature review

2.1 Face Anti-Spoofing

Starting from first principles, we have that an instance of biometric data is named a biometric. A biometric spoof is an artificial mimic of a real biometric. In those terms, Anti-Spoofing is then understood as a technical measure against spoofing [8]. Clearly, data depiciting, representing or symbolizing faces of real world people is a biometric. Therefore, it rationally follows to elaborate upon the family of concepts that fall under Face Anti-Spoofing (FAS). Figure 2.1 shows an example of such system.

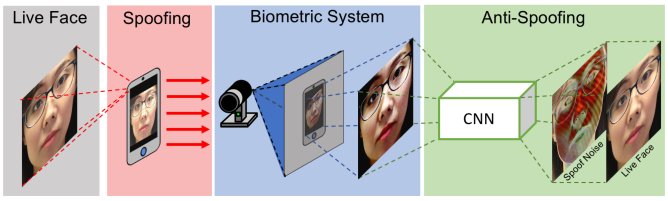


Figure 2.1: Example of a neural network based FAS system [1]

2.2 Biometric Privacy-Enhancing Techniques

Techniques that try to mitigate or alleviate privacy concerns over biometric data by concealing information that can allow to identify individuals while leaving everything else unchanged are usually called Biometric Privacy-Enhancing Techniques (B-PETs) [9] Applying this techniques has a drawback reduced biometric utility. As such, B-PETs strive to reach a balance between

privacy and biometric utility.

2.2.1 Characteristics

There are different criteria for classifying B-PETs. They are usually grouped according to (i) kind of biometric data (ii) guarantees about they privacy enhancement (thresholds, bounds, biometric utility completely retained under X scenario, etc.) (iii) possibility of reconstructing the original data after applying a technique (iv) whether the information is concealed to human and/or machines (v) whether a technique is applied to the data itself, a distilled representation of it (embeddings, templates, etc.) and/or systems that manipulate the data. [9]

As there is a plethora of criteria, an algorithmic taxonomy for B-PETs has been proposed [9], which is constructed by making the broadest categorization according to the level some B-PETs operate. These levels are (i) image, (ii) representation and (iii) inference A brief summary [9] of each is given:

Image-level techniques

At this level biometric data in the form of still images and video is manipulated with enhancing *visual privacy* (privacy from other humans) as a goal. Synthesis, obfuscation and adversarial approaches all fall under this approach.

Representation-level techniques

In some cases access to *raw* data might not be possible and the so-called templates, sets of represensative features distilled from other biometric data such as images, are used instead. A particular technique that should be highlighted due to its prominent guarantees is homomorphic encryption, which in some scenarios allows for complete biometric utility retention.

Inference-level techniques

The stage at which biometric data (enhanced or not) is matched or classified against other biometric templates is called *inference*. Machines and not humans are the "public" to which data is concealed from. The added value from these techniques comes from two sources (i) ensuring

that privacy is enhanced since conception and (ii) enhancing systems that did not have privacy as a goal in mind.

This paper will be solely concerned with *image-level* techniques.

2.2.2 Evaluation

Three criteria are generally used to evaluate B-PETs: i) *efficiency*, ii) biometric utility *retention* and iii) *robustness* against reversal attacks. A description of each criterion along with an adjunt metric follows:

1. *Efficiency* Efficiency of B-PETs is usually evaluated with two data-sets, one with privacy enhanced data and the other without. In the case of verification scenarios, efficiency is measured according to how well a system recognises privacy enhanced data in comparison to unenhanced data; that is, determining if the system is still capable of tying privacy enhanced data to a correct identity. Efforts have been made to quantify privacy enhancenment. For example, a metric named *privacy gain* is defined as follows:

$$\mathrm{PG} = (1 - R_p) - (1 - R_o) \tag{2.1}$$

   Where R is recognition performance and p denotes the results in the enhanced data meanwhile o does it for the original set.

2. *Biometric Utility Retention* Since providing a balance between privacy enhancement and biometric utility is one of the purposes of B-PETs, it is natural to expect that a metric measuring to what extent a technique fulfills that was expected. However, it should be noted that *utility* is highly application-dependent. Nonetheless, a concrete metric for measuring biometric utility retention has been used in past research [9]. Assume x and y are two distinct images:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.2}$$

   Where:

   - μ_x and μ_y are the average (mean) values of the input images x and y respectively.
   - σ_x and σ_y are the standard deviations of x and y respectively.

- σ_{xy} is the covariance of x and y.
- C_1 and C_2 are constants to stabilize the division with a weak denominator.

SSIM stands for Structure Similarity Index Measure, representing the degree of structural similarity between x and y.

2.3 Related work

2.3.1 Data augmentation methods

Variety and amount of data more often than not accounts for a large part of the performance of a model. There is no exception for FAS systems. Since face image of real-world people is often difficult to gather in comprehensive and extensive means due to privacy concerns, ethical diversity of and time-effort issues, automated solutions ought to be found. In fact, previous work [11] demonstrates that data augmentation with improvements in recognition performance are possible via adversarial methods and diffusion. On another hand, these methods can also be considered as B-PETs since they fall under the image-level synthesis family of techniques. As such, those models are susceptible of (a) being bypassed by representation level techniques and (b) consuming conspicous amounts of computational resources [9]. While the former drawback can not be addressed by the techniques yet to be explored in this paper, the latter can be alleviated (or so does the author think).

Chapter 3

Objectives

- Apply biometric privacy-enhancing techniques to datasets of images.
- Fine-tune face anti-spoofing models with privacy-enhanced images.
- Compare the fine-tuned models according to efficiency, biometric utility and robustness.
- Evaluate the balance between privacy enhancement and performance degradation according to the fine-tuned and original models.

Chapter 4

Methods

4.1 Biometric Privacy-Enhancing Techniques Selection

Given the computationally simple nature of image-level techniques some of these were selected to enhance that served as fine-tuning and evaluation material to the Face Anti-Spoofing. Concretely, three techniques were chosen: Artistic Style Transfer (AST), Gaussian Blur, and Pixelation. A short introduction and demonstration of each follows:

4.1.1 Artistic Style Transfer

Some specific architectures of neural networks have allowed to reinterpret the images within an specific artistic style [4]. An example of this technique can be seen in 4.1.

(a) Original

(b) Privacy-Enhanced

Figure 4.1: Effect of Artistic Style Transfer

It must be said that previous work on B-PETs doesn't acknowledge this technique as orthodox. This thesis will attempt to assess its validity as a legitimate B-PET.

4.1.2 Gaussian Blur

This technique applies a filter that smooths images in a non-linear way. An example can be seen in 4.2. Mathematically speaking, this technique is parametrised by two parameters: a kernel size and a parameter named sigma x.

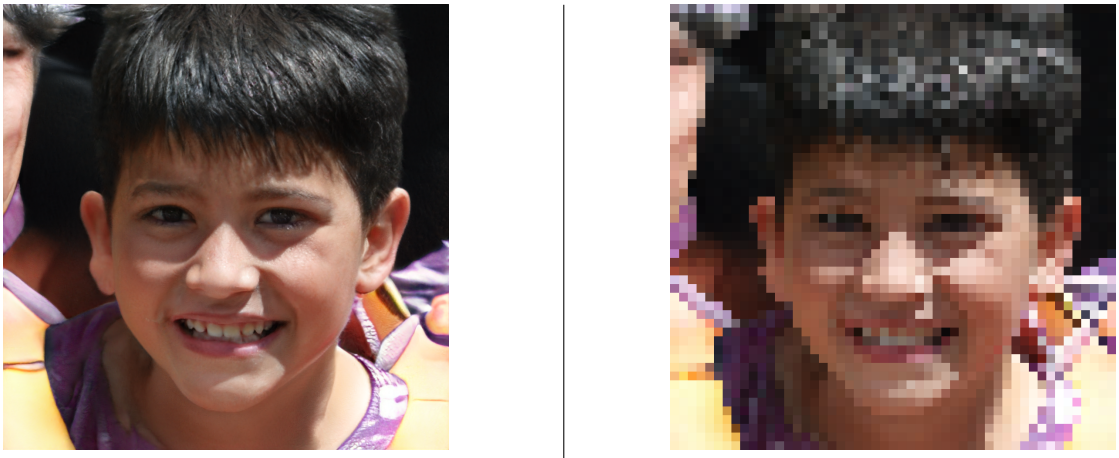
(a) Original

(b) Privacy-Enhanced

Figure 4.2: Effect of Gaussian Blur

4.1.3 Pixelation

This technique is rather straightforward. It takes an image and joins near-by pixels according to an specified block size. An example can be seen in 4.3. This technique is parametrised by the parameter block size.



(a) Original

(b) Privacy-Enhanced

Figure 4.3: Effect of Pixelation

4.1.4 Criteria

After selecting the techniques, these were measured according to the Privacy Gain and Biometric Utility metrics mentioned in 2.2.2, Privacy-Gain (PG) and Structure Similarity Index Measure (SSIM) within the Yale Faces [5] dataset. 4000 images of the dataset were randomly selected, to derive results faster.

Since the Privacy-Gain metric measures how effective a B-PET is at concealing identities of people, an identity classifier of the Yale Faces dataset had to be trained. In particular, transfer-learning with the aid of the EfficientNetB3 architecture was performed. As the model is only expected to classify images from the Yale Faces dataset, no train-test split had to done.

On another hand, both Gaussian Blur and Pixelation are parametrised so these techniques were subjected to an hyperparameter search. Since AST is computationally expensive only one configuration was used, one that was determined by the artistic style of Wassily Kandinsky.

Per each configuration the PG and SSIM metrics were calculated. The criterion for selecting the best configuration of each technique was finding a balance between privacy-gain and biometric utility. To do so, the contribution of the metrics was equally weighted i.e. multiplying each metric by $\frac{1}{2}$ and then summing them. This variable was named "Score".

See the images from 4.4 to 4.9. According to the images the best technique for Gaussian Blur is determined by a kernel size of 91 and for Pixelation it is the one determined by a block size of 20.

4.2 Fine-Tuning of Face Anti-Spoofing Architectures

In order to fine-tune the three Face Anti-Spoofing architectures, images in the privacy-enhanced domains had to arranged. All of the architectures are fed with images and output a prediction, which is attributed as to whether the model "thinks" an image is authentic or spoofed.

Per each selected B-PET, 7500 images of the Yale Faces dataset and 7500 from the Style GAN dataset [7] were enhanced. The former were labeled as authentic and the latter as spoofs, which is methodologically sound according to previous research. [11]

Furthermore, some of these models rely on the detection of faces in the input images, either

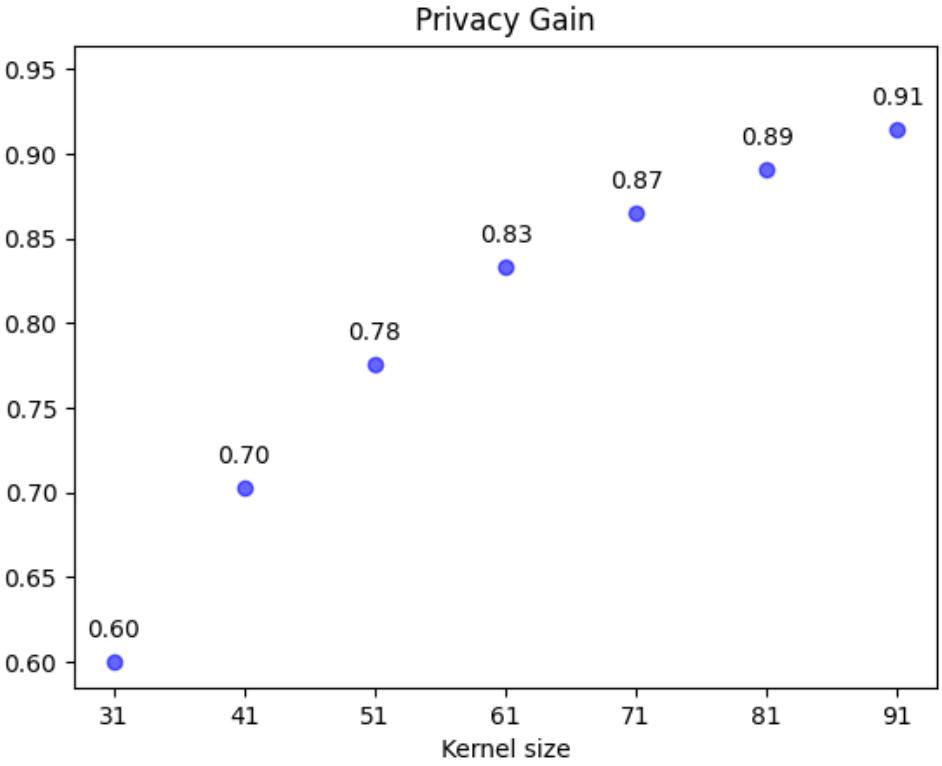


Figure 4.4: Privacy Gain in respect to Kernel size for Gaussian Blur

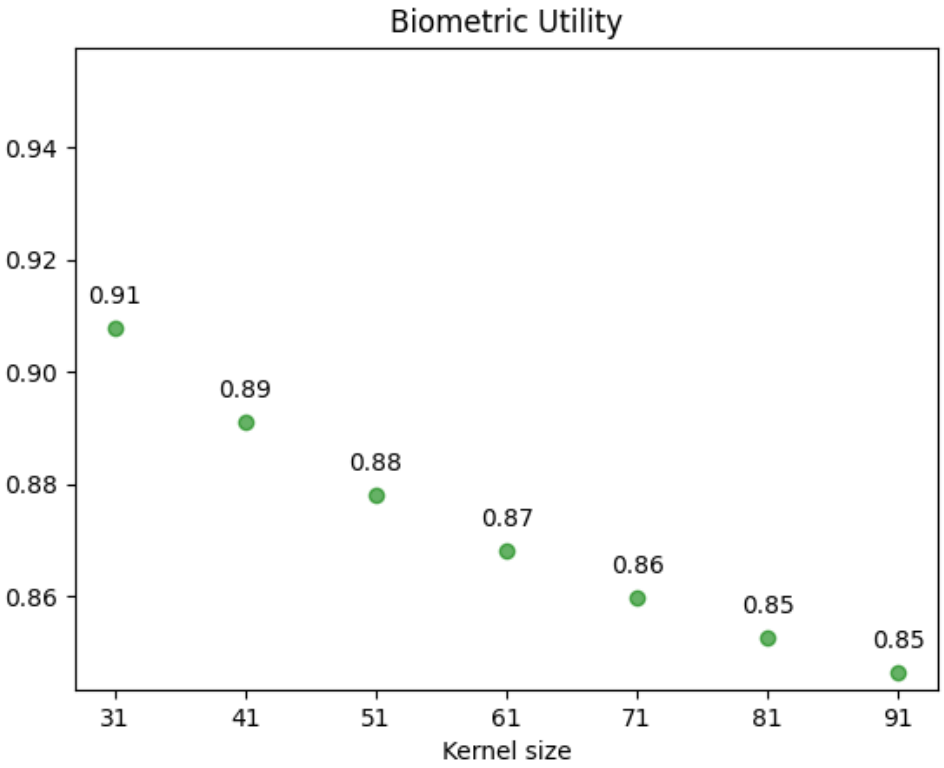


Figure 4.5: Biometric Utility in respect to Kernel size for Gaussian Blur

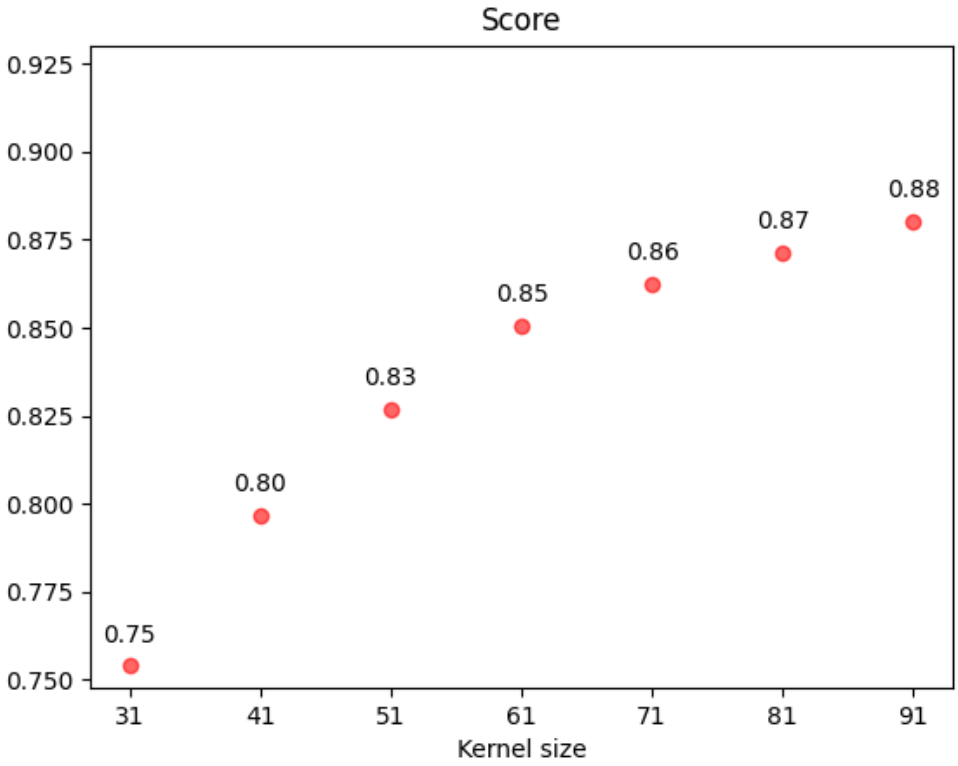


Figure 4.6: Score in respect to Kernel size for Gaussian Blur

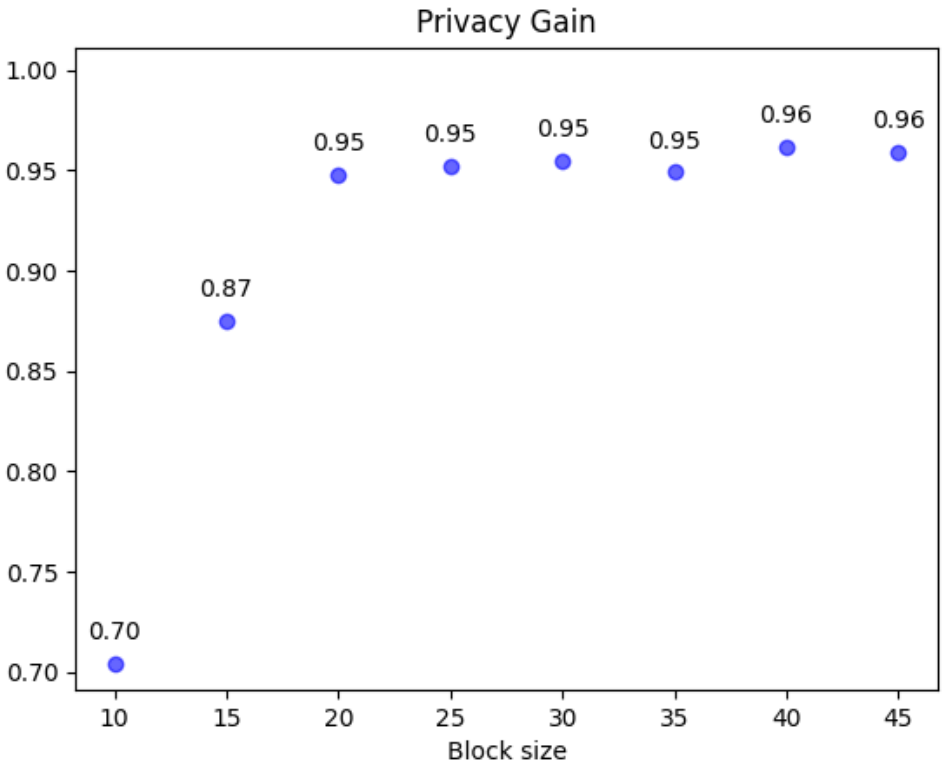


Figure 4.7: Privacy Gain in respect to Block size for Pixelation

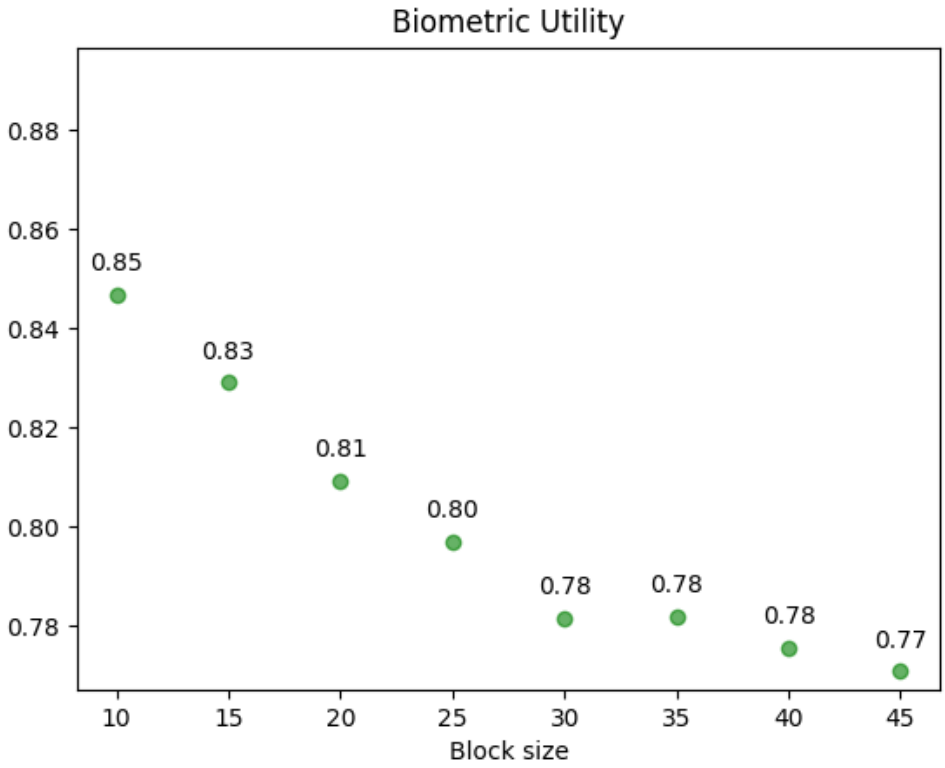


Figure 4.8: Biometric Utility in respect to Block size for Pixelation

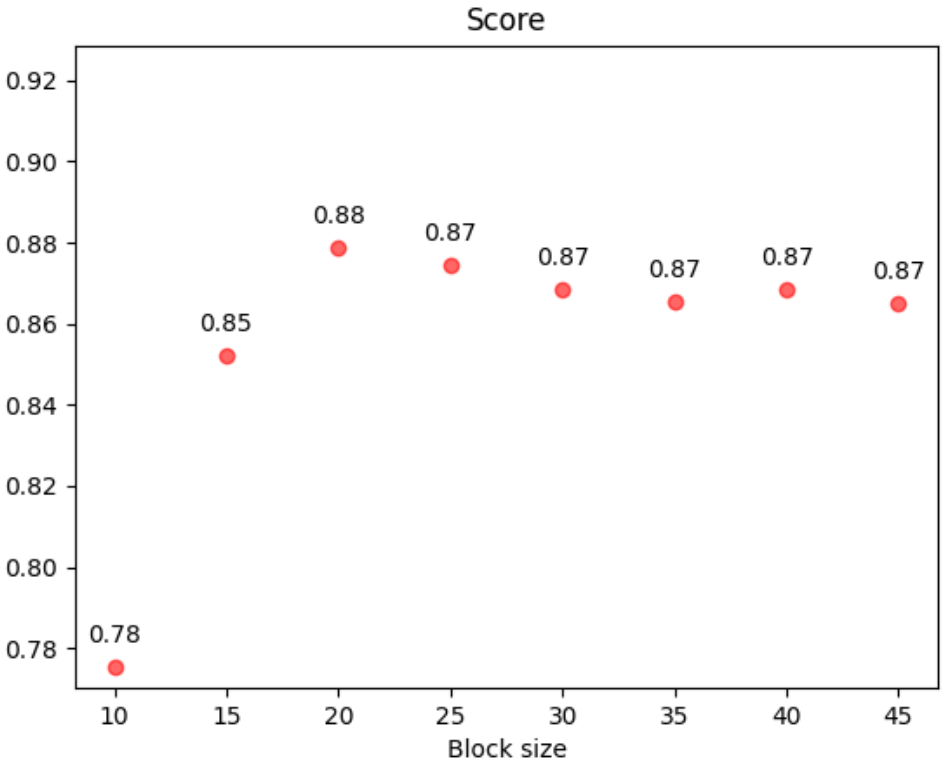


Figure 4.9: Score in respect to Block size for Pixelation

in the training or evaluation phase. Since it is expected that the biometric utility of privacy-enhanced images is reduced in comparison to "normal" images, face detectors in the privacy-enhanced domains had to be trained. These detectors were trained with the Labeled Faces in the Wild (LFW) dataset [6], a 15K-wide collection of face images specifically tailored for deep-face recognition. The bounding dimensions of the faces found in the dataset were calculated with the unmodified dataset. Then, the dataset was enhanced with each one of the B-PETs and fed into a deep-regression model that outputs the bounding-dimensions of faces in the privacy-enhanced domains. The deep-regression model wasn't built from scratch but rather built on top of ImageNetV2 and repurposed for the face-detection task i.e. transfer-learning.

On another hand, these models were previously fine-tuned on a broader, more comprehensive datasets in order to enhance their dexterity at recognising spoofed images [11]. As such, the fine-tuning process of this experiment will be performed on top of the aforementioned models.

Each FAS model was fine-tuned with each of the selected B-PETs. Since there are three models and three B-PETs, **9 models were obtained**.

An overview of each model's architecture and differentiating traits will be provided:

4.2.1 Silent Face Anti-Spoofing (SFAS)

This model's architecture is comprised by a standard convolutional neural network and a Fourier spectrum convolutional neural network. The former receives images, processes them and at some point it outputs a feature map that is concomitantly fed The former receives images, processes them and at some point it outputs a feature map that is concomitantly fed to the latter, which calculates the Fourier spectrum of the latter and then is pipelined to three convolutional layers. The output of the model itself is determined by the first network. However, the training phase is driven by a loss metric that equally weighs the losses of the standard and Fourier spectrum networks. [3]

The main motivation for adding the Fourier Transform to the recognition pipeline is rooted on the intuition that the spectra of real images is fundamentally distinct from the spectra of spoofs. For example, abrupt changes in texture and brightness - ubiquitous in spoofs - are detected by Fourier spectra analysis.

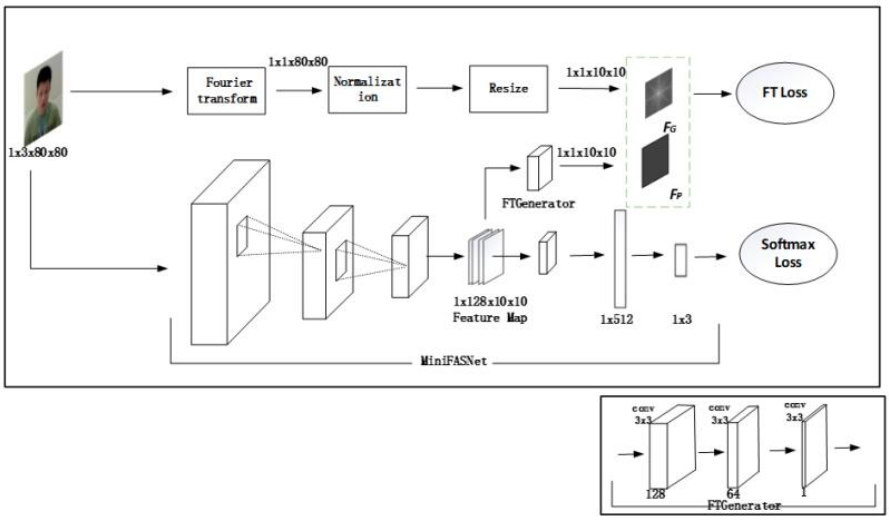


Figure 4.10: Architecture of SFAS

4.2.2 Liveness Detection Face Anti-Spoofing (LFAS)

The architecture of this model is rather straightforward since it is a deep convolutional neural network with no relevant innovation or addition unlike the other two models [2]. It's architecture can be detailed at 4.11

4.2.3 Object Face Anti-Spoofing (OFAS)

In contraposition to the previous two convolutional networks, this architecture relies on the brightness of images as a proxy for authenticity. On a more technical level, histograms of colors in YCrCb and LUV color spaces are calculated, meshed together in a row vector and then fed into an ensemble of tree classifiers who give the final prediction [11]. It's architecture can be detailed at 4.12.

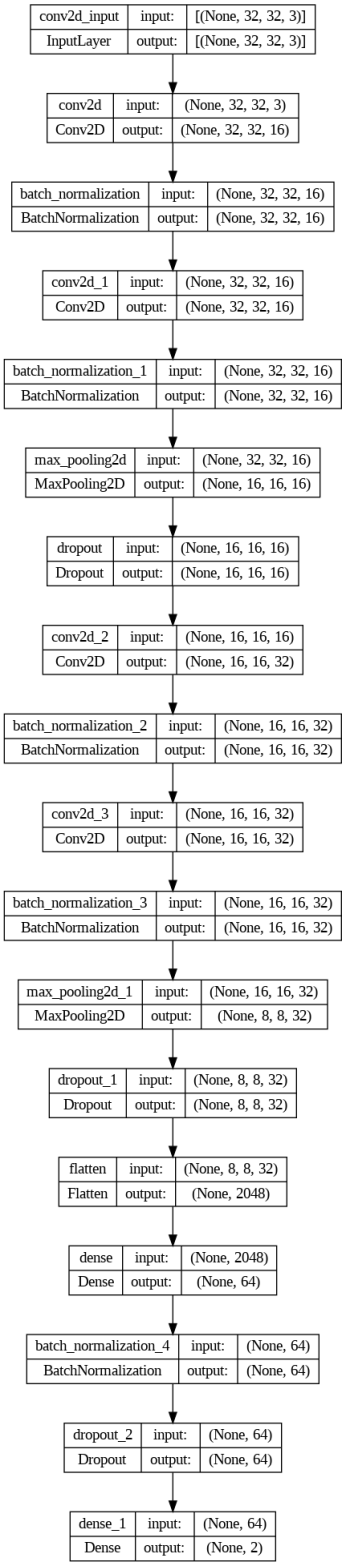


Figure 4.11: The convoluted architecture of LFAS

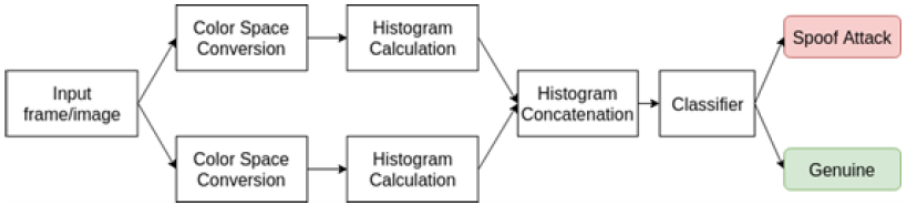


Figure 4.12: Architecture of OFAS

Chapter 5

Evaluation

5.1 Dataset

In order better simulate the conditions of Face Anti-Spoofing production systems, the MSU-MFSD dataset [12] was employed. Such dataset contains 280 video recordings of genuine and attack scenarios from 35 different subjects. This dataset is completely orthogonal to all the other datasets that were used to choose the most fit B-PETs and aid the fine-tuning process Vega carried out in its research, meaning that all of the metrics considered in the evaluation process constitute as reliable proxies of how the privacy-enhanced FAS would fare in real-world conditions. 20 frames were evenly sampled from all the videos, **constituing then the test dataset with roughly 5600 images**.

5.2 Evaluation scheme

The proposed evaluation scheme goes as follows: consider an architecture X and a B-PET Y. There will be an incoming stream of images which will be transformed with Y. Then, the images will be fed to two models: the baseline model of X and the architecture of X that was privacy-enhanced with Y. **This adds up to 18 evaluation runs**. The idea of adopting this evaluation scheme is to gauge (i) to what extent a technique Y degrades the performance of the models and (ii) how does each model X respond to being fine-tuned with B-PETs.

Abbreviation	**Meaning**
TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives

Table 5.1: Abbreviations for terms used in posterior metrics

Metric	**Formula**
Accuracy	$\frac{TP + TN}{\text{Total Population}}$
Precision	$\frac{TP}{TP + FP}$
APCER	$\frac{FP}{TN + FP}$
NPCER	$\frac{FN}{FN + TP}$
ACER	$\frac{\text{APCER} + \text{NPCER}}{2}$

Table 5.2: Set of metrics to calculate & analyse

5.3 Metrics

In order to make the comparison of the distinct models and B-PETs, the following metrics will be taken into account:

To provide more context, Attack Presentation Classification Error Rate (APCER) stands for how often an spoof is classified as authentic by a FAS model, Normal Presentation Classification Error Rate (NPCER) stands for how often a genuine image is classified as fake and Average Classification Error Rate is balanced measure between the two. These are metrics specifically tailored for the comparison and benchmarking of FAS systems.

Chapter 6

Results

Given the rich mosaic between techniques and models, it is of interest to watch the impact of B-PETs from the a technique-centric and model-centric viewpoints.

6.1 Performance by technique

Figures from 6.1 to 6.3 detail the impact in a technique-basis.

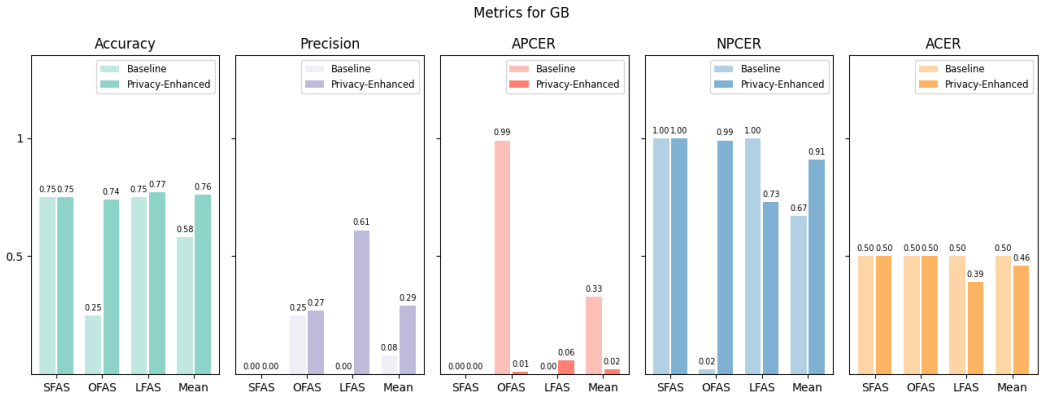


Figure 6.1: Metrics for Gaussian Blur (GB)

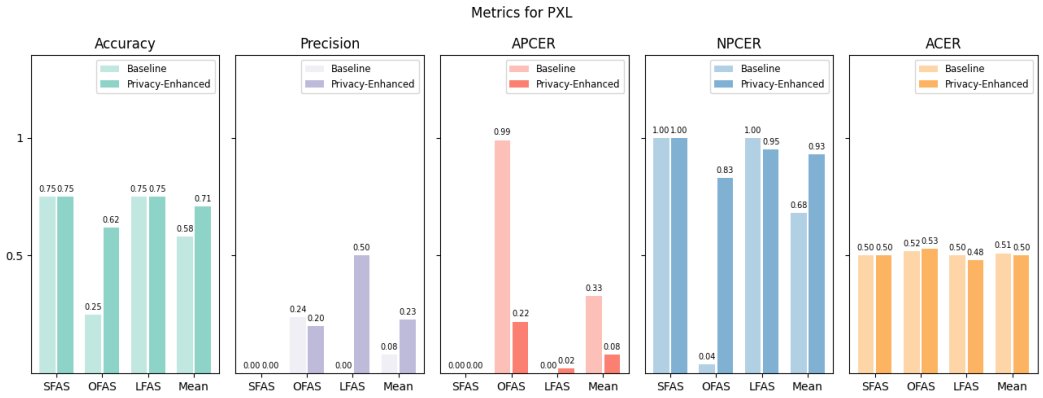


Figure 6.2: Metrics for Pixelation (PXL)

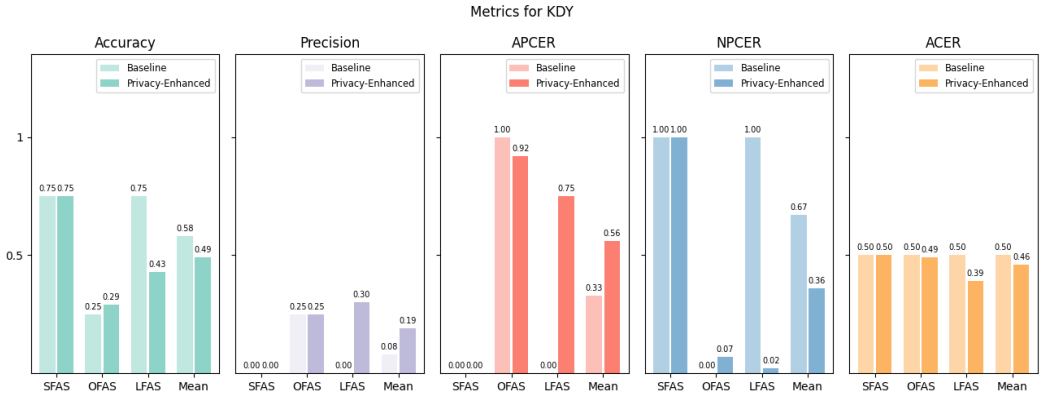


Figure 6.3: Metrics for Artistic Style Transfer (AST)

6.2 Performance by model

Figure from 6.4 to 6.6 detail the impact in a model-basis.

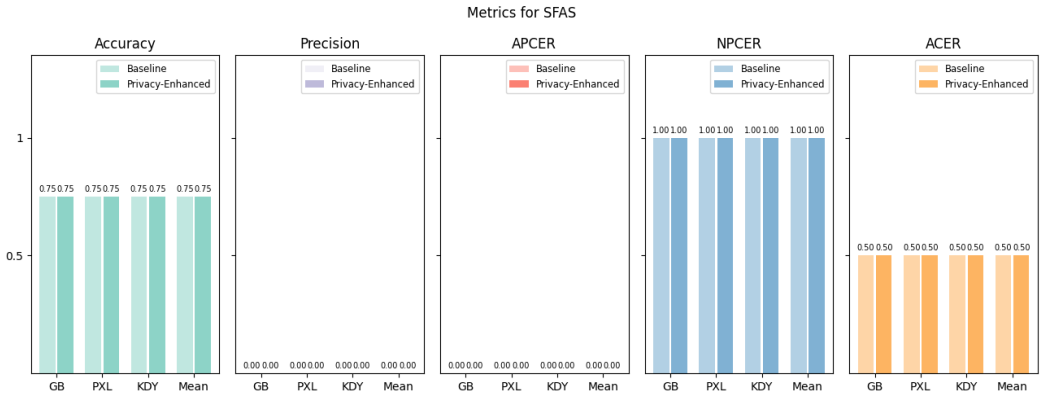


Figure 6.4: Metrics for SFAS

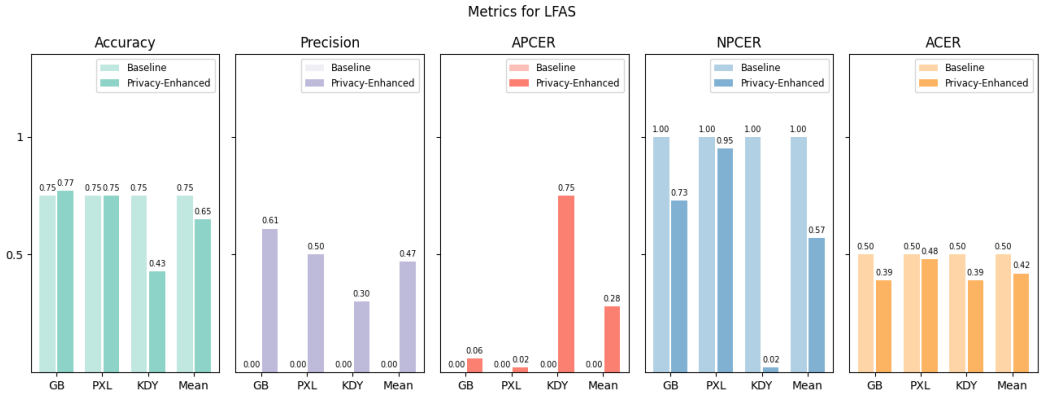


Figure 6.5: Metrics for LFAS

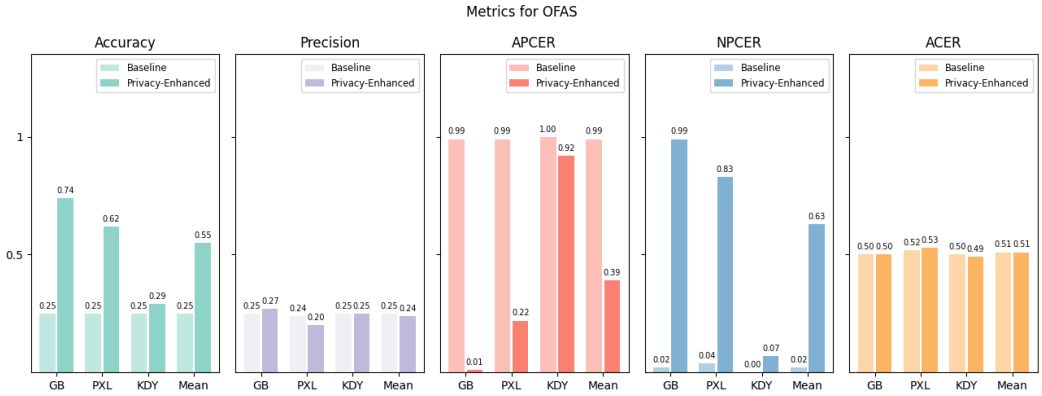


Figure 6.6: Metrics for OFAS

Chapter 7

Discussion

Technique-wise, we can see that according to the ACER metric - which is a balanced measure of both APCER and NPCER - no technique confers a relevant difference when the transformed filter of images is subject to FAS detection. This means that fine-tuning with privacy-enhanced images is not needed and efforts could be directed in qualitatively distinct directions.

Architecturally-wise, it can be boldly asserted that fine-tuning with the privacy-enhanced images considerably degrades the performance of the models, as it can be seen with the extreme example of SFAS.

Overall, given that image-level techniques are computationally simple to setup and evaluate, they constitute a decent first line of defense in terms of privacy. However, the results of this workpiece suggest that image-level techniques do not constitute an *effective* way of introducing privacy mechanisms to FAS models. This conclusion is further nourished by the fact qualitatively distinct B-PETs and FAS models were employed.

Chapter 8

Further work

Further developing the conclusions outlined in the previous section, it can be confidently asserted that image-level techniques by themselves are not enough to provide inference-time privacy guarantees in the deployment of production-ready FAS systems. Nonetheless, some lines of research worth exploring would be the following:

- Adding an intermediate representation of biometric live-footage and apply representation-level techniques to them.
- Mixing image-level techniques. An example would be combining the Artistic Style Transfer and Gaussian Blur techniques and then analyse its usefulness with the approach followed in this work.
- Provided both approaches above don't yield experimentally interesting results, apply differential privacy frameworks, although the author doesn't think this leads to particularly relevant outcomes for inference-time privacy guarantees, that is.

Bibliography

The references are sorted alphabetically by first author.

[1] Face anti-spoofing, face presentation attack detection. `https://cvlab.cse.msu.edu/project-face-anti.html`, 2022.

[2] Face liveness detection anti-spoofing web app. `https://github.com/birdowl21/Face-Liveness-Detection-Anti-Spoofing-Web-App`, 2022.

[3] Minivision AI. Silent face anti-spoofing. `https://github.com/minivision-ai/Silent-Face-Anti-Spoofing`, 2020.

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015.

[5] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001. DOI 10.1109/34.927464.

[6] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[8] Stan Z. Li. *Encyclopedia of Biometrics*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 0387730028.

[9] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J. Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacyenhancing face biometrics: A comprehensive survey. *Trans. Info. For. Sec.*, 16:4147–4183, jan 2021. ISSN 1556-6013. DOI 10.1109/TIFS.2021.3096024. URL `https://doi.org/10.1109/TIFS.2021.3096024`.

[10] Beate Roessler and Judith DeCew. Privacy. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[11] C. Vega Fernández. Estrategias para la generación sintética de imágenes y su aplicación a escenarios de aumentación de datos en el desarrollo de sistemas face anti-spoofing, 2023.

[12] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. DOI 10.1109/TIFS.2015.2400395.