General Requirements on Synthetic Fingerprint Images for Biometric Authentication and Forensic Investigations

Andrey Makrushin
Otto von Guericke University
Magdeburg, Germany
andrey.makrushin@ovgu.de

Christof Kauba
Department of Computer Sciences,
University of Salzburg
Salzburg, Austria
ckauba@cs.sbg.ac.at

Simon Kirchgasser
Department of Computer Sciences,
University of Salzburg
Salzburg, Austria
skirch@cs.sbg.ac.at

Stefan Seidlitz
Otto von Guericke University
Magdeburg, Germany
stefan.seidlitz@ovgu.de

Christian Kraetzer
Otto von Guericke University
Magdeburg, Germany
christian.kraetzer@iti.cs.
uni-magdeburg.de

Andreas Uhl
Department of Computer Sciences,
University of Salzburg
Salzburg, Austria
uhl@cs.sbg.ac.at

Jana Dittmann
Otto von Guericke University
Magdeburg, Germany
jana.dittmann@ovgu.de

ABSTRACT

Generation of synthetic biometric samples such as, for instance, fingerprint images gains more and more importance especially in view of recent cross-border regulations on security of private data. The reason is that biometric data is designated in recent regulations such as the EU GDPR as a special category of private data, making sharing datasets of biometric samples hardly possible even for research purposes. The usage of fingerprint images in forensic research faces the same challenge. The replacement of real datasets by synthetic datasets is the most advantageous straightforward solution which bears, however, the risk of generating "unrealistic" samples or "unrealistic distributions" of samples which may visually appear realistic. Despite numerous efforts to generate high-quality fingerprints, there is still no common agreement on how to define "high-quality" and how to validate that generated samples are realistic enough. Here, we propose general requirements on synthetic biometric samples (that are also applicable for fingerprint images used in forensic application scenarios) together with formal metrics to validate whether the requirements are fulfilled. Validation of our proposed requirements enables establishing the quality of a generative model (informed evaluation) or even the quality of a dataset of generated samples (blind evaluation). Moreover, we demonstrate in an example how our proposed evaluation concept can be applied to a comparison of real and synthetic datasets aiming at revealing if the synthetic samples exhibit significantly different properties as compared to real ones.

IH&MMSec '21, June 22–25, 2021, Virtual Event, Belgium.

ACM ISBN 978-1-4503-8295-3/21/06.
https://doi.org/10.1145/3437880.3460410

CCS CONCEPTS

• **Security and privacy** → **Biometrics**;

KEYWORDS

Biometrics; Synthetic Samples; Fingerprints

ACM Reference Format:
Andrey Makrushin, Christof Kauba, Simon Kirchgasser, Stefan Seidlitz, Christian Kraetzer, Andreas Uhl, and Jana Dittmann. 2021. General Requirements on Synthetic Fingerprint Images for Biometric Authentication and Forensic Investigations. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '21), June 22–25, 2021, Virtual Event, Belgium.* ACM, Innsbruck, Austria, 12 pages. https://doi.org/10.1145/3437880.3460410

1 INTRODUCTION

Recent cross-border regulations on security of private data impose severe restrictions on preserving, sharing and processing of person related data. Since biometric data such as face, iris and fingerprint images are considered a special category of private data, sharing databases of biometric samples has become very difficult. Moreover, biometric samples may include privacy sensitive attributes such as gender, age, ethnicity, or health status [37]. These facts hinder both scientific research and industrial development of reliable biometric access control systems and forensic investigation tools.

Gathering large-scale datasets of biometric samples has been always an issue. Even disregarding that acquiring biometric samples is very time- and resources-intensive, it is also tremendously hard to gather unbiased datasets of sufficient scale.

In general, there are two ways to compile a dataset of biometric samples that are not linked to individuals: (i) gather biometric samples from real persons and anonymize/de-identify them, (ii) artificially generate so-called synthetic biometric samples. The first option is clearly disadvantageous due to the need for data gathering as well as for tedious removing of privacy sensitive attributes.

The second option seems to be superior from all points of view, nonetheless bearing the risk of generating "unrealistic" samples or "unrealistic distributions" of samples which may visually appear realistic.

In this paper we endeavor answering the question: What is the quality of synthetic biometric samples, or more precisely, which requirements should the synthesized datasets meet to replace datasets of real samples. After defining hard and soft requirements, we suggest how to validate whether the requirements are fulfilled or how to adapt the sample generation process to fulfill most of the requirements.

The paper is organized as follows: An overview of fingerprint generation approaches is given in Section 2 (see Table 1). As our main contribution, in Section 3, we propose requirements on synthetic datasets along with validation metrics (see Table 2). Section 4 comprises evaluation of several exemplary datasets regarding some of the proposed requirements. In Section 5, we draw our conclusions and present future work.

2 RELATED WORKS

There exist two concepts of generating samples for biometric or forensic research: model-based and data-driven. Early works were rather focused on model-based generation. For fingerprints, the process is an inverse to the extraction of minutiae. Starting from a set of minutiae, a fingerprint area, an orientation map and a frequency map are estimated. Then, the iterative pattern growing approach draws ridges along the orientation lines by applying Gabor filters [7]. This approach has been implemented in the software tool SFinGe [6] and then reimplemented in the software tool Anguli[1]. Model-based fingerprint texture synthesis is also discussed in [20]. The Anguli generator has been used to create the "groundtruth" fingerprint images[2] for a set of subsequently degraded fingerprint samples used in a fingerprint impainting and denoising competition [1, 9].

The general problem of model-based generation approaches is lack of realism [28]. In particular, the limitations can be summarized as follows:
(1) Independent generation of ridge orientation and minutiae models leads to the fact that minutiae distributions could be generated without having a valid ridge orientation field;
(2) Ridge-lines as well as spacing between ridge-lines has a constant width making the detection of synthetic patterns an easy task. High accuracy while discriminating real and synthetic fingerprints was demonstrated in [8];
(3) Using masterprints for ridge-line pattern, generation may lead to non-consistent ridge flow leading to unrealistic fingerprint images;
(4) Non-considering local minutiae configurations may lead to unrealistic minutiae configurations. This drawback was exploited in [13] to reliably discriminate real and synthetic fingerprints.

In contrast to modeling, the modern trend is a data-driven generation of realistic fingerprint images. Thanks to recent development of artificial neural networks and especially Variational Autoencoders (VAE) and Genarative Adversarial Networks (GAN) very impressive results have been demonstrated not only with images of human faces but also with fingerprint images.

The first effort to synthesize fingerprints using a Wasserstein GAN is made in [4] aiming at generating so-called master fingerprints that match multiple original fingerprints. Later on, in [27], a connectivity imposed GAN is introduced and applied to two datasets: FVC-2006 and PolyU. In [2], fingerprints are synthesized by a variational autoencoder. In [10], a lightweight GAN is proposed for creating 128x128 pixel fingerprint patches and compared with five established GAN architectures based on 64x64 pixel patches. The size of generated images in the aforementioned publications is rather insufficient. In [5], a combination of an autoencoder and an adapted Wasserstein GAN is used for synthesizing 512x512 pixel fingerprint patches. The realistic appearance of synthetic fingerprints is dramatically improved in [38] by applying CycleGAN to transfer texture from real fingerprints to conventionally synthesized ridge-line patterns with added sweet pores. An alternative approach for generating high-resolution realistic fingerprints is proposed in [33]. Here, GAN is combined with a super-resolution network. The next breakthrough is done in [28] by incorporating identity information into the fingerprint synthesis network which is based once again on combining autoencoder and Wasserstein GAN.

Despite numerous efforts to generate high-quality fingerprints, there is still no common agreement on how to define "high-quality" and how to prove that the generated samples are realistic enough. Moreover, datasets of synthetic fingerprints are compiled for different purposes. The characteristics of synthetic fingerprints that are strictly required for validation of Automated Fingerprint Identification Systems (AFIS) might not be necessary for analyzing latent fingerprints in forensic investigations and vice versa. In [34], for instance, the focus is on generation of latent fingerprints. The NIST SD 27 dataset is used as a source of training samples and the software tool StirTrace[3] is applied for training data augmentation. StirTrace simulates artifacts that are identified in the literature [16–18, 26] as relevant for forensic analysis of latent fingerprints at a crime scene. In particular, StirTrace is used for simulation of sensor-characteristic artifacts represented by Random additive noise and Random Gaussian noise, sensor scan artifacts represented by salt and pepper noise as well as different acquisition conditions artifacts represented by rotation, re-scaling and cropping of a fingerprint image.

GAN-based generation of latent fingerprints with pre-defined characteristics can be made possible if a training dataset is augmented by patterns representing environmental conditions, substrate influence, features of an acquisition device and acquisition conditions, and such a dataset is used for training of a GAN architecture which implements disentanglement learning. The study in [34] is not focused on conditional generation but introduces the concept which is capable of such functionality.

Another important aspect is the simulation of fingerprints partially corrupted due to errors in the acquisition process [21] or skin diseases [22]. To the best of our knowledge there still exist no GAN models capable of doing it.

[1] https://dsl.cds.iisc.ac.in/projects/Anguli/
[2] http://chalearnlap.cvc.uab.es/dataset/32/description/
[3] https://sourceforge.net/projects/stirtrace/

Table 1: State of the art in GAN-based generation of fingerprint images together with the requirements to synthetic data and metrics applied

Paper	Generation approach	Source DB	Max. image dim.	Addressed requirements (see section 3.1)
Minaee et al. '18 [27]	Finger-GAN: Connectivity imposed GAN	FVC'06, PolyU	64x64	R+D (Frechet Inception Distance to training samples)
Bontrager et al. '18 [4]	DeepMaster-Prints: WassersteinGAN	NIST SD9, FingerPass DB7	128x128	R (Visual inspection), A (Matching against training samples)
Cao & Jain '18 [5]	Autoencoder + WassersteinGAN	NIST SD27, NIST SD14	512x512	R (NFIQ2, minutiae configuration), D (Pair-wise matching), A (Matching against training samples)
Attia et al. '19 [2]	Variational Autoencoder	no info	no info	no info
Fahim et al. '20 [10]	Lightweight GAN	LivDet	256x256	R (Visual inspection), D (MS-SSIM scores)
Riazi et al. '20 [33]	SynFi: GAN + super-resolution DCNN	NIST SD9	256x256	R (Visual inspection + fake detection)
Mistry et al. '20 [28]	Convolutional Autoencoder + WassersteinGAN + Identity Prior	Longitudinal fingerprint records	512x512	R (NFIQ2, minutiae configuration), D (Pair-wise matching), A (Matching against NIST SD4 samples)
Wyzykowski et al. '20 [38]	Anguli + Pores + Style Transfer using CycleGAN + super-resolution CNN	PolyU HRF	512x512	R (Visual inspection, FP recognition test, Human perception test)
Seidlitz et al. '21 [34]	PGGAN, StyleGAN StyleGAN2	NIST SD27	512x512	R (NFIQ2), A (Matching against training samples)

Table 1 summarizes the state of the art in GAN-based fingerprint generation together with the requirements on the synthetic data and the metrics applied. First and the foremost concern of all studies is the realistic appearance, which is measured either objectively by the NFIQ2 metric [31] or statistics describing minutiae configuration or subjectively by a visual inspection. Many studies also evaluate diversity and anonymity of synthetic samples. Diversity is usually measured by checking an "impostor" distribution or to be more precise by cross-matching of all generated samples and comparison of the resulting matching score histograms to those of real fingerprint datasets. Anonymity is provided if there are no matches between generated and training samples. Hence, each generated sample is matched against all samples used for GAN training. Fingerprint matching is conducted by applying commercial of-the-shelf (COTS) solutions (e.g. Neurotechnology VeriFinger SDK[4] or Innovatrics Fingerprint Recognition SDK[5] or by open source software (e.g. NIST/Bozorth3[6]). All in all, we can conclude that still there is no consensus on how to evaluate the quality of a generative model or even the quality of a dataset of generated samples.

3 THEORETICAL CONCEPT

As our main contribution, we propose requirements on synthetic biometric data generated by a generative model. These requirements are defined in Section 3.1 and summarized in Table 2. They are classified into hard (technical) and soft (ethical) requirements. For each requirement we also propose a validation method and a generator adaptation technique.

In addition to general considerations, we adapt our proposed requirements to synthetic fingerprints. Moreover, we distinguish between two practically relevant but distinct fingerprint application scenarios: biometric authentication and forensic investigation of latent fingerprints bearing in mind that the requirements may have different importance.

3.1 Definition of the Requirements

The requirement ***Realistic appearance*** **(R)** can be interpreted differently. In our considerations realistic appearance encompasses two criteria: 1. the naked eye cannot tell apart real and generated fingerprints; 2. ridge-line patterns appear statistically natural. The first one is rather important for manual investigations of fingerprints e.g. training of forensic officers. The second one is essential for reliable AFIS validation. The appearance to the naked eye is subjective and can be tested only in a human experiment. In contrast, statistical characteristics (e.g. number of cores and deltas, frequency of ridge lines, amount of minutiae, background noise) can be validated by comparing distributions of these characteristics in real and synthetic data.

For fingerprints, the recently proposed NFIQ2 metric should perfectly cover the second criterion. Initially, the NFIQ metric has been designed to check whether fingerprints acquired by commercial biometric sensors are good enough to be further used in AFIS for reliable person identification. The advanced metric NFIQ2 is optimized to assess the quality of 500 dpi live-scanned fingerprints coming from adults and captured with optical scanners, but can be easily adapted to assess the quality of other types of fingerprints e.g. digitalized latent fingerprints [3]. The generation of fingerprints with a realistic appearance is a challenge for both model-based and data-driven generation approaches. Moreover, it is stated that the existing model-based approaches generally lack realism [28].

The requirement ***Sufficiently high image resolution*** **(I)** is not an issue for model-based generation. For data-driven generation, in contrast, image resolution is of major importance because the size of input and output images is an inherent property of generative

[4]https://www.neurotechnology.com/verifinger.html
[5]https://www.innovatrics.com/biometrics-for-oem-solutions/fingerprint-recognition/
[6]https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis

Table 2: General requirements on synthetic biometric data

Requirement	Validation	Generator adaptation
	Hard (technical) requirements	
Realistic appearance (R)	Subjective test, data-specific quality metric	Selection of a suitable generative model
Sufficiently high image resolution (I)	Trivial check of image dimensions as well as dimensions of a fingerprint region	Turn the input/output image size from an inherent property to a tunable parameter
Data anonymity (A)	Differential privacy validation by matching of synthetic and training samples	Randomization of identity related attributes in training
Diversity/Uniqueness (D)	Biometric cross-comparison of generated samples and comparison of score distributions with those of real fingerprint datasets	A-posteriori uniqueness validation or pseudo-random generation
Controllable generation (C): same subject vs. different subjects	Comparison of intra-class and inter-class distributions of matching scores to each other	Integration of an identity prior into the training process and conditional generation
Reflection of basic characteristics of ground truth (training) data (B)	Comparison of distributions of certain characteristics (real vs. synthetic), Frechet Inception Score (FIS)	Selection of a suitable generative model, Integration of fingerprint characteristics into the training process and conditional generation
	Soft (ethical) requirements	
Equal distribution of privacy-related attributes (E), such as gender, age, ethnicity etc.	Detector-based validation	Integration of privacy-related attributes into the training process and conditional generation

The **application of these requirements is done in this paper in** the following sections:

Realistic appearance (R)	In sections 3.4 for an exemplary application of the requirements for a structured assessment of existing generative models and in 4.2.1 for assessing our own experiment
Sufficiently high image resolution (I)	In section 3.4
Data anonymity (A)	In section 3.4
Diversity/Uniqueness (D)	In section 4.2.2
Controllable generation (C):	In section 4.2.2
Reflection of basic characteristics (B)	Reserved for future work
Equal distribution of privacy-related attr. (E)	Reserved for future work

neural networks. The image size in pixels should be high enough to cover the whole area of a fingerprint captured at scanning resolution of at least 500 dpi which is required by FBI-compliant systems [19]. Speaking of a fingerprint area that is approximately 1.0x0.7 inch, the image size should exceed 500x350 pixels. The minimal resolution to work with sweat pores is 1000 dpi, requiring the image size of at least 1000x700 pixels. Generation of such images is a hard competition for e.g. GAN-based generation approaches. The probably most elegant solution is the Progressive Growing GAN [23] where the image size increases during the learning process gradually covering more and more details of an image.

The requirement ***Data anonymity*** **(A)** means the anonymity of generated fingerprints and implies that no person in the world possess these particular fingerprints. In this formulation, the requirement cannot be fulfilled because there is no dataset of all possible real fingerprints and a synthetic fingerprint could match an existing real one just by chance. The more modest requirement is that a synthetic fingerprint does not belong to any of subjects who provided their fingerprints for a training dataset used for fingerprint generation. Note that for model-based generation approaches where the minutia map is generated completely randomly, the data anonymity requirement is fulfilled by definition (just because of non-existence of the training dataset). For data-driven generation approaches, data anonymity is closely connected with the term 'differential privacy'. A loose definition would be: 'an algorithm is differentially private if an observer cannot extract any information about a particular subject by analyzing algorithm's output.' Differential privacy in application to GAN-generated data is addressed in [36].

Validation of differential privacy of a fingerprint generator could happen by matching each synthetic fingerprint against all real fingerprints used for generator training, looking for exactly one match. If a synthetic fingerprint matches k ($k > 1$) original instances, then k-anonymity is indicated. For fingerprints, it is highly likely that by mixing images the GAN generator strongly modifies the minutia maps of the training fingerprints so that the minutia maps in synthetic fingerprints have a random nature. However, there is no guarantee that all synthetic fingerprints do not match original ones. The ratio of anonymous fingerprints in the whole number of generated fingerprints should indicate the 'anonymity score' of a generative model.

In order to improve the anonymity score of a generative model, one should either switch to conditional generation by including an identity prior to the generator training or randomly change the minutia maps of training samples.

The requirement ***Diversity/Uniqueness*** **(D)** implies that the generated fingerprint images are significantly different from each other. Due to the mode collapse often encountered in GAN training, many generated samples may be very similar. Checking whether

the synthetic samples are diverse enough could happen by cross-comparison of them, meaning that each generated sample is matched against all other generated samples. Ideally, the distribution of the cross-comparison scores should not strongly deviate from that of real fingerprints. However, a general comparison between a dataset of synthetic fingerprints and real fingerprints is hardly possible because the distribution of cross-comparison scores may strongly vary from one dataset of real fingerprints to another.

Compilation of diverse synthetic fingerprint datasets is possible by including an a-posteriori uniqueness validation meaning that each generated fingerprint is matched against all previously generated fingerprints and preserved in a dataset only in case of no single match. The other option is, as in the case of data anonymity, to rely on conditional generation by including an identity prior to the generator training. So, every new fingerprint can be generated using a unique identity (minutia map). Note that for model-based approaches the diversity issue is easily solved by randomization of model parameters (basic pattern, minutia map) during the generation.

The requirement ***Controllable generation*** **(C)** implies the ability of a generative model to generate not only random inter-class samples but also intra-class samples which is very important for a generative model. For model-based fingerprint generation the solution is trivial - keep basic pattern, minutia map and sweat pores locations and randomly change all other model parameters. For data-driven generation, the standard solution is disentanglement learning and further conditional generation [29] meaning that either class labels or minutia maps should be integrated into the training process.

The validation whether the network is able to generate samples of the same class could happen by comparison of intra-class and inter-class distributions of matching scores to each other.

The requirement ***Reflection of basic characteristics of ground truth data*** **(B)** implies that generated fingerprints preserve inherent characteristics of real fingerprints used for generation. Since model-based generation requires no training samples, this requirement is solely relevant for data-driven generation. If we can clearly define and extract a certain characteristic from an image and, moreover, describe its presence by a scalar value, the direct comparison of the value distribution in real and synthetic datasets is a trivial and straightforward solution. In a more general case, when the presence of a characteristic cannot be formalized and represented by a single value, the comparison of distributions by Frechet Inception Distance [15] is the only objective way to validate whether the characteristics of real samples have been transferred to synthetic samples. Similarly to controllable generation, the standard solution here is disentanglement learning and further conditional generation, however, focusing not on identity of a fingerprint but on some characteristic of training samples.

In fact, data-driven fingerprint generation has a huge potential for targeted fabrication which might be of interest for biometric authentication (e.g. simulating different biometric fingerprint sensor types such as capacitive or thermal sensors) and is of special interest for forensic investigations. In [30], a wide range of characteristics have shown to be relevant. They concern all stages along the corresponding signal processing and pattern recognition pipeline. An non-exhaustive list would contain for example: a) pipeline selection: the type of fingerprint under consideration (rolled vs. plain vs. latent fingerprints), fingerprint development method impact (esp. for latent fingerprints); b) sample/signal acquisition: sensor and sensor scan characteristics, smudgy fingerprints (and other trace characteristics such as fibers or other traces overlapping a latent print) and substrate characteristics, different acquisition conditions; c) pre-processing: handling of application scenario specific artifact characteristics (e.g. for printed fingerprint contexts or to compensate arbitrary image sizes). This non-exhaustive list could easily be extended and has seen in the past the emergence of dedicated image post-processing tools like StirTrace (see Section 2) for data augmentation and derivation of datasets with simulated artifacts for forensic purposes, including the training of dactyloscopic experts.

The requirement ***Equal distribution of gender, age, ethnicity, etc*** **(E)** is important to avoid biases in evaluation datasets which often lead to a wrong picture of reality. Training with biased data leads to biased models which can be discriminatory or even harmful to humans. Since fingerprints may preserve privacy-related attributes (other than the minutia map), it is required that samples with each expression of each attribute are equally well presented in a dataset. This requirement to a generative model is currently rather theoretical due to the lack of a precise picture on how exactly personal characteristics such as gender, age or ethnicity manifest themselves in fingerprint patterns.

Technically, the adaptation of a generative model to include the control of privacy-related attributes is similar as for generating certain fingerprint characteristics on demand.

3.2 Evaluation perspective

Our proposed requirements are initially formulated for evaluation of generative models. However, the same instrument can be applied to evaluate datasets of synthetic biometric samples to be used in biometric or forensic evaluations. In the case of evaluating a generative model, we possess two datasets: a set of real samples used for training and a set of generated samples. In the case of evaluating a dataset of generated samples, real samples used for training are not known. This is why we call the first evaluation case *informed* (or generic evaluation) and the second case - *blind* evaluation. While the informed evaluation comprises all formulated requirements, the blind evaluation is usually limited to the following subset: realistic appearance (R), sufficient image resolution (I) and diversity (D). However, if a probe dataset includes several instances of one finger, the requirement controllable generation (C) should also be validated. If a probe dataset includes metadata describing characteristics of the generated sample such as, for instance, "simulation of fingerprints scanned by a particular optical sensor", then the validation of the requirement (B) could also make sense.

3.3 Importance of the requirements for Biometrics and Forensics

Depending on the application scenario for which the synthetic biometric samples are generated, the requirements discussed above

receive different weights regarding their importance. If the focus is on evaluation of biometric authentication systems, it is highly important that a generative model is able to generate both intra-class and inter-class samples which are on the one hand sufficiently diverse (D) and on the other hand reflect the characteristics of real samples especially the sensor type influence (the requirements C and B).

In contrast, if the focus is on scenarios involving the manual analysis of fingerprints, like in forensic crime scene investigation, the most important requirements are realistic appearance (R), sufficiently high image resolution (I) and the ability to reproduce the basic characteristics of real samples like a fingerprints aging effect or a substrate influence (B).

The requirements realistic appearance (R), data anonymity (A), diversity (D) and equal distribution of privacy-related attributes (E) are important independently of the field in which the dataset of synthetic fingerprints is going to be applied.

Besides requirement engineering aspects (i.e. specifying characteristics that should be addressed in data generation for a specific application scenarios), the requirements introduced in this paper can also be used to provide a structured assessment of both: a data synthesis process (see Section 3.4) and a resulting dataset (see Section 4). As an additional benefit, using this set of requirements as a projection surface allows for a better comparison between generative models (or datasets).

3.4 Exemplary application of the requirements for a structured assessment of existing generative models

An example of evaluation and (less formal) comparison of several generative models can be found in our previous study [34]. We tested three GAN architectures (Progressive Growing GAN [23], StyleGAN [24] and StyleGAN2 [25]) for generation of latent fingerprints trained using around 40k fingerprint patches sampled from the NIST SD 27 dataset. Prior to training, the set of fingerprint patches was augmented by their filtered versions increasing the number of training samples to approximately 500k. Since the generative models are trained iteratively and each snapshot of a model can be used for generation of fingerprint images, the evaluation was conducted for only one snapshot of each model. The number of training iterations is proportional to the parameter kImages referring to the number of real images presented to the network during the training phase. Usually, the parameter kImages is set prior to training and the training lasts until the number of iterations is reached. However, due to time limitations, training can be stopped earlier without lose of the generative model performance. Aiming at fingerprint image quality assessment during the training, some snapshots were used for generation of 1000 fingerprint images and estimating the average NFIQ2 scores of them. For each of the three GAN architectures one snapshot was picked for the evaluation so that the balance between the training time and the average fingerprint image quality in terms of the NFIQ2 metric was held. Hence, the generative models are evaluated by assessing a set of 1000 fingerprint images generated by the aforementioned model snapshots from random seeds. The main assumption is that 1000 fingerprint images selected in such a way reflect the general ability of the Generator network to produce images of a certain quality. The fingerprint images generated by the latter snapshots should in average possess the same properties.

Our previous study in [34] explicitly addresses only three requirements: realistic appearance (R), data anonymity (A) and sufficiently high image resolution (I). The remaining requirements diversity (D), conditional generation (C), reflection of basic characteristics of training data (B) and equal distribution of privacy-related attributes (E) are either addressed implicitly or not addressed at all.

The more structured informed evaluation of the generative models from [34] is done in the following:

- *Realistic appearance* (R): A *subjective* visual inspection of the fingerprint images suggests that all three generative models are capable of producing synthetic fingerprints which cannot be told apart from real fingerprints by the naked eye. An *objective* evaluation on the realistic appearance is done by calculating the NFIQ2 scores and comparing the histograms of NFIQ2 scores of real and synthetic fingerprints. It is shown that synthetic fingerprint patches generated by any of the three considered generative models have on average slightly higher NFIQ2 scores than the real fingerprint patches used for training. The reason for this phenomenon is discussed in detail in [34]. This trivial analysis of NFIQ2 scores enables us to conclude that the considered generative models are capable of generating realistically appearing fingerprint patches. The NFIQ2 scores of 266 out of 1000 images generated with the Progressive Growing GAN, 371 out of 1000 images generated with the StyleGAN and 191 out of 1000 images generated with the StyleGAN2 exceed the threshold of 35 meaning the very good to excellent quality of fingerprints.
- *Sufficiently high image resolution* (I): The requirement for high image resolution is automatically fulfilled by the selection of a certain GAN architecture and a proper snapshot of a generative model that generates fingerprint patches with the size of 512x512 pixels at a resolution of 1000 ppi.
- *Data anonymity* (A): Since the evaluation of a data anonymity requires matching of fingerprints, the fingerprint images should be of sufficient quality (see requirements R and I) for reliable extraction of minutiae. Hence, the data anonymity requirement is checked only with the fingerprint images for which a NFIQ2 score is higher than 35. Each of those synthetic fingerprints is matched against all real fingerprints used for training of the corresponding generative model. The matching is done by means of NIST tools: MINDTCT for minutiae extraction and Bozorth3 for comparison of minutiae lists. The threshold for the Bozorth3 score has been set to 30, meaning that if exactly one Bozorth3 score is higher than 30, the generated fingerprint is reported to have a match in the training dataset and, therefore, cannot be considered anonymous. The number of reported anonymous fingerprints are 169 out of 266 (63.53%) for the Progressive Growing GAN, 182 out of 371 (49.06%) for the StyleGAN and 106 out of 191 (55.50%) for the StyleGAN2. Using our terminology, the ratios of anonymous fingerprints, namely 63.53%, 49.06% and 55.50%, can be seen as estimations of the "anonymity score" of the corresponding generative models.

- *Diversity* (D): For each of the three generative models, we visually inspected the 1000 generated fingerprint patches and concluded that all patches appear significantly different to each other. However, no objective validation has been conducted.
- *Controllable generation* (C): Due to the absence of fingerprint identity information during the training, all three models trained are not capable of generating several different fingerprint images for one fingertip.
- *Reflection of basic characteristics of training data* (B): The visual inspection conducted in [34] suggests that the generated images have very similar characteristics to the NIST SD27 patches, however no formal objective validation has been conducted.
- *Equal distribution of privacy-related attributes* (E): Since the source dataset includes no information about privacy-related attributes in training fingerprint images and the fact that there is no consensus on how particular privacy-related attributes manifest themselves in fingerprint patterns, validation of this requirement is not possible yet.

4 EVALUATION

An application of our evaluation concept is demonstrated in an example of blind evaluation of two synthetic fingerprint datasets generated by SFinGe (FVC2002 Db4 and FVC2004 DB4).
The main goal of the following evaluation is to assess the postulated requirements Realistic appearance (R), Diversity (D), and Controllable generation (C) for these datasets. This is done by verifying whether it is possible to differentiate synthetic and real fingerprint images by comparing their quality value and comparison score distributions. This analysis is refined by the application of a statistical test based on the results of histogram metrics utilized to measure the overlap of the considered distributions.

4.1 Evaluation methodology

4.1.1 *Datasets.*

PLUS FP Ageing: For the acquisition of this database, 10 different fingerprint capturing devices have been used. Five of them are optical (including two multispectral scanners), four are capacitive (one swipe) and one is thermal. In total there are 108106 fingerprint samples, acquired in 4 sessions over a time-span of 2 years from 59 different subjects (50 of these subjects participated in all 4 capturing sessions). Only a subset of the database was used, containing samples from the third session for six selected capturing devices, which are: RealScan G1 (optical), URU5100 (optical), Lumidigm V311 (optical multispectral), Lumidigm M311 (optical multispectral), IB Columbo (capacitive) and NB-3010-U (thermal).

CASIA FP Ageing: The "CASIA Fingerprint Subject Ageing Version 1.0" is a publicly available dataset focusing on biometric template ageing effects[7]. The 5880 samples were collected from 49 different subjects in two sessions with 4 years timespan in between. The imprints' acquisition was done with three different capturing devices, two of them are optical (Digital Persona URU4000 and URU4500), and the last one, the Atmel TouchChip TCRU1C scanner, is capacitive.

FVC2002 and FVC2004: These benchmark fingerprint datasets were used during the 2002 and 2004's fingerprint verification contest. Both contain 4 subsets, 3 of them captured by real fingerprint scanners while the fourth one contains synthetically generated samples by the SFinGe generator [6]. Each subset includes 800 fingerprint samples from 100 fingers, 8 imprints per finger.

- FVC2002: Db1 - optical sensor "TouchView II" by Identix, Db2 - optical sensor "FX2000" by Biometrika, Db3 - capacitive sensor "100 SC" by Precise Biometrics and Db4 - synthetic generated samples using SFinGe v2.51
- FVC2004: DB1 - optical sensor "V300" by CrossMatch, DB2 - optical sensor "U.are.U 4000" by Digital Persona, DB3 - thermal sweeping sensor "FingerChip FCD4B14CB" by Atmel and DB4 - synthetic fingerprint samples using SFinGe v3.00

4.1.2 *Tools implementing metrics.*

Two different kinds of metrics were used, the first one to evaluate realistic appearance and the second one to evaluate diversity.
For fingerprints, a natural way to assess the realistic appearance (requirement R) is to employ standard fingerprint quality metrics. Hence, we used NFIQ2[8], NFIQ1 [35] and several quality metrics suggested by Olsen et al. [32]:

- Frequency Domain Analysis (FDA): This metric is based on extracting the ridge-valley signature of the imprints in a block-based manner by computing the Discrete Fourier Transformation of the signature to determine the frequency of the sinusoid following the ridge-valley structure. This results in a local quality score, which is averaged over all blocks to derive the final quality value.
- Gabor quality (GAB): It operates on a per-pixel basis by calculating the standard deviation of responses from a Gabor filter bank. This filter bank describes areas in the imprint by measuring the regular ridge-valley pattern and thus, there will be a high response from filters containing fingerprint information, while in areas containing background or unclear ridge-valley structure the Gabor response of all orientations will be low and constant.
- Gabor-Shen quality (GSH): This block-based metric also makes use of a Gabor filter bank. The filter response of each Gabor kernel is computed on the pixels in each block and a standard deviation is computed on the responses. Using thresholding, each block is determined to be either foreground or background and also of poor or good quality.

All those metrics have a different output range with higher values indicating better quality.

In order to assess the diversity (requirement D), we performed cross-comparison experiments of the samples using several state-of-the-art fingerprint recognition systems: Neurotechnology VeriFinger 11.0 SDK[9] (output score range [0, 2000]) and Innovatrics

[7] http://biometrics.idealtest.org/dbDetailForUser.do?id=15

[8] https://github.com/usnistgov/NFIQ2

[9] https://www.neurotechnology.com

ANSI/ISO SDK[10] (output score range [0, 1000]) being two representatives of minutiae-based fingerprint recognition systems as well as Fingercode (FC) and Phase Only Correlation (POC) [14] (output score range [0, 1]) as non-minutiae based ones.

In the first case, we generate quality scores for each sample and, in the second case, comparison scores between pairs of mated fingerprints (genuine scores) as well as pairs of non-mated fingerprints (impostor scores). Based on these scores, histograms and fitted density distributions (to these histograms) are utilised to further evaluate/illustrate the results. In order to go beyond visual inspection of distribution plots and to quantify the outcome, the histograms were compared using the following standard histogram metrics: Chi-Squared (Chi), Histogram Intersection (HI) and Kullback-Leibler (KL) [12]. All three metrics exhibit an output range of [0, 1] on our data, where 0 means highest possible correspondence between the histograms (for Kullback-Leibler) or lowest possible correspondence (for Chi-Squared and Histogram Intersection). Note that e.g. for Chi-Squared the numerical values have been inverted as compared to the original definition to achieve a homogeneous interpretation of the values in the employed MatLab software[11]. The computed comparison and quality score density distributions were fitted using a uni-variate kernel density estimate based fitting.

A five-times repeated random sampling of quality or comparison scores was done to ensure an equal number of scores for each dataset. For imposter scores each time 1000 scores were sampled, while for genuine scores and the quality values 600 were selected. This became necessary as the number of samples in the datasets varies to a great extent. All plots presented in the following show the averaged results from the random samples independently drawn for each dataset (the corresponding standard deviation is always lower than 1%).

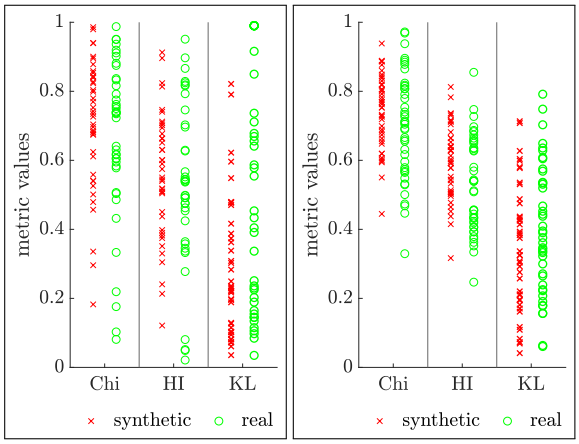

Figure 2: Plots showing the high overlap of histogram comparison metric results for GSH (left) and NFIQ2 (right).

[10] https://www.innovatrics.com
[11] "Histogram distances" https://de.mathworks.com/matlabcentral/fileexchange/39275-histogram-distances

4.2 Validation of requirements

4.2.1 Realistic appearance (R).

Each of the applied metrics (NFIQ2, NFIQ1, FDA, GSH and GAB) highly differ from each other in describing different imprint's characteristics. Thus, the overall trends with regards to synthetic and real fingerprint samples are highly varying. Figure 1 shows some examples of these variations detected in the quality value distributions of FDA (top left), GSH (top right), GAB (bottom right) and NFIQ2 (bottom left). Not all of the evaluated datasets are depicted, only those which exhibit a clearly different behaviour to the remaining ones. The results obtained using NFIQ1 exhibit a similar trend to GAB. The utilised line style and colour for each dataset are the same for all the following score distribution plots (see Figure 4). The choice of the synthetic dataset, either FVC2002 Db4a or FVC2004 DB4A does have hardly any influence on the plots and thus, the same trend can be observed regardless of which FVC synthetic subset is taken into account. This can be seen if Figure 3 is compared with the top left image of Figure 1. Figure 3 depicts also the FDA quality results, but using FVC2004-DB4A, while the top left plot of Figure 1 was generated using FVC2002-Db4a sample information. As the overall tendency for both synthetic datasets are the same we decided to show only FVC2002-Db4a plots in all following figures.

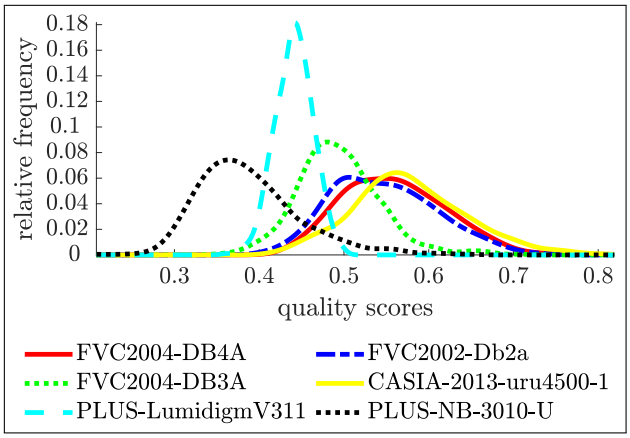

Figure 3: FDA quality value distributions using FVC2004 synthetic samples.

Several oddities can be observed. First, in the top left (FDA) at least one real sample dataset seems to perfectly overlap the distribution of the synthetic samples. Other distributions exhibit a reduced overlap, lowest for both PLUS datasets and there is no overlap present for GAB and GSH. This indicates that the FVC synthetic data might have been adjusted to exhibit similar quality levels as the FVC real data according to NFIQ and other ridge structure based quality metrics (like FDA).

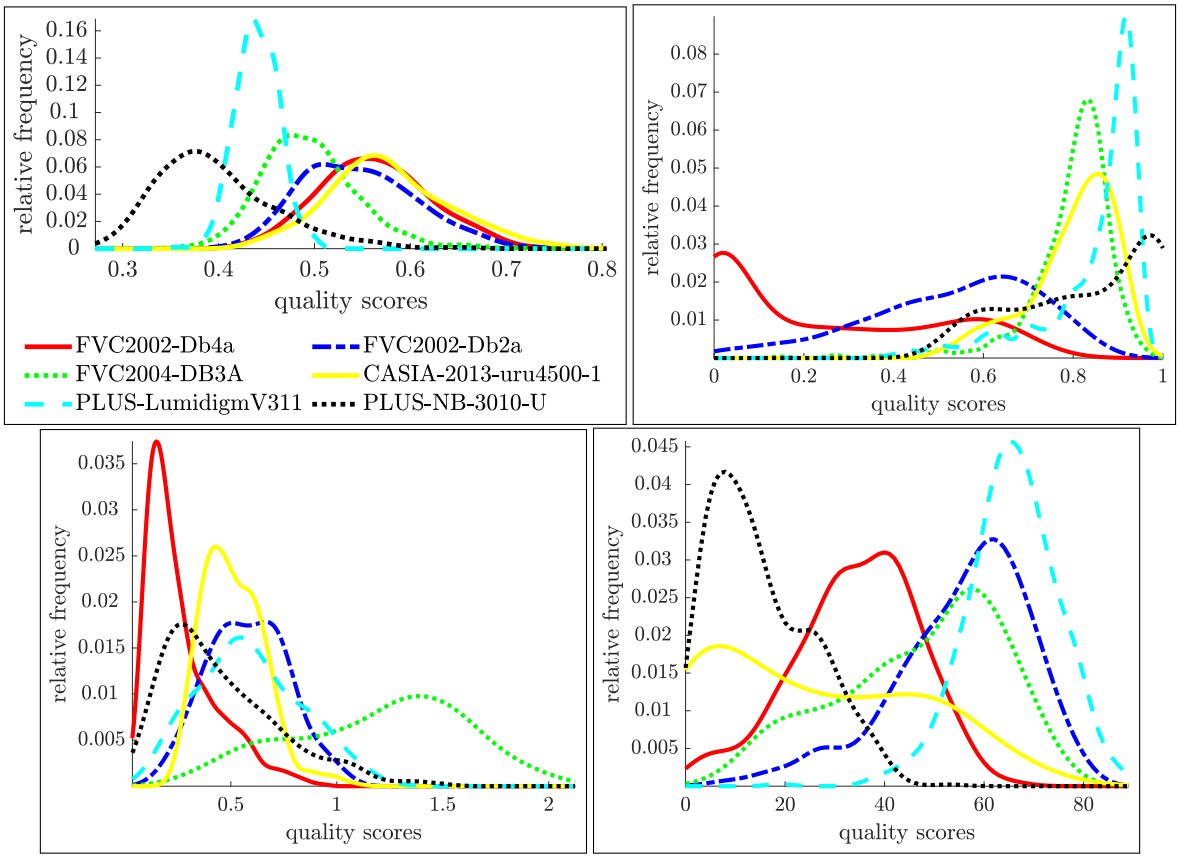

Figure 1: Quality value distributions of samples from FVC, CASIA and PLUS datasets using FDA (top left), GSH (top right), GAB (bottom left) and NFIQ2 (bottom right). The line colour, line style and naming of top left plot was used in all following distribution plots.

Table 3: P-values of Mann-Whitney-U tests of histogram metric values from real and synthetic quality value distribution comparisons. The null hypothesis in each case is defined such that the histogram metric values of real and synthetic score distribution comparisons are from continuous distributions with equal medians with an alpha significance level of 0.01**.**

	NFIQ2	NFIQ	FDA	GAB	GSH
Chi	0.04	0.02	0.62	0.79	0.06
HI	0.04	0.05	0.64	0.89	0.09
KL	0.73	0.98	0.65	0.83	0.0002

In the top right example, showing GSH distributions, the synthetic distributions can be characterised as skewed to the left, while all (depicted) real datasets are skewed to the right. However, as the positive/negative skewness is relatively flat, a high overlap of the quality values can still be detected from a visual point of view.

Overall, the fingerprint quality distributions of synthetic and real data are sometimes similar (for specific sensor choices), often very different (for other sensors), but not clearly separable. Hence, a clear assessment of requirement (R) is not possible based on this visual inspection.

In order to approach a more quantitative statement, we apply histogram metrics to compare the distributions. The aforementioned visual overlap also affects the utilised histogram metrics to a large extent, which is shown in Figure 2, visualising the results obtained by Chi, HI and KL when: a) metric values between synthetic and real sample distributions are marked as red crosses and b) metric values between real and real sample distributions are marked as green circles. In both cases, 21 values have been randomly sampled from all computed histogram comparisons. In contrast to the distributions visualised in Figures 1 and 4, all the evaluated datasets have been considered for the random sampling. For case a) only results involving FVC2002 Db4a samples are shown (results for FVC2004 DB4A would be highly similar). For case b) (real vs real samples) all combinations of real data distribution comparisons were computed. The number of included metric values is balanced for both cases.

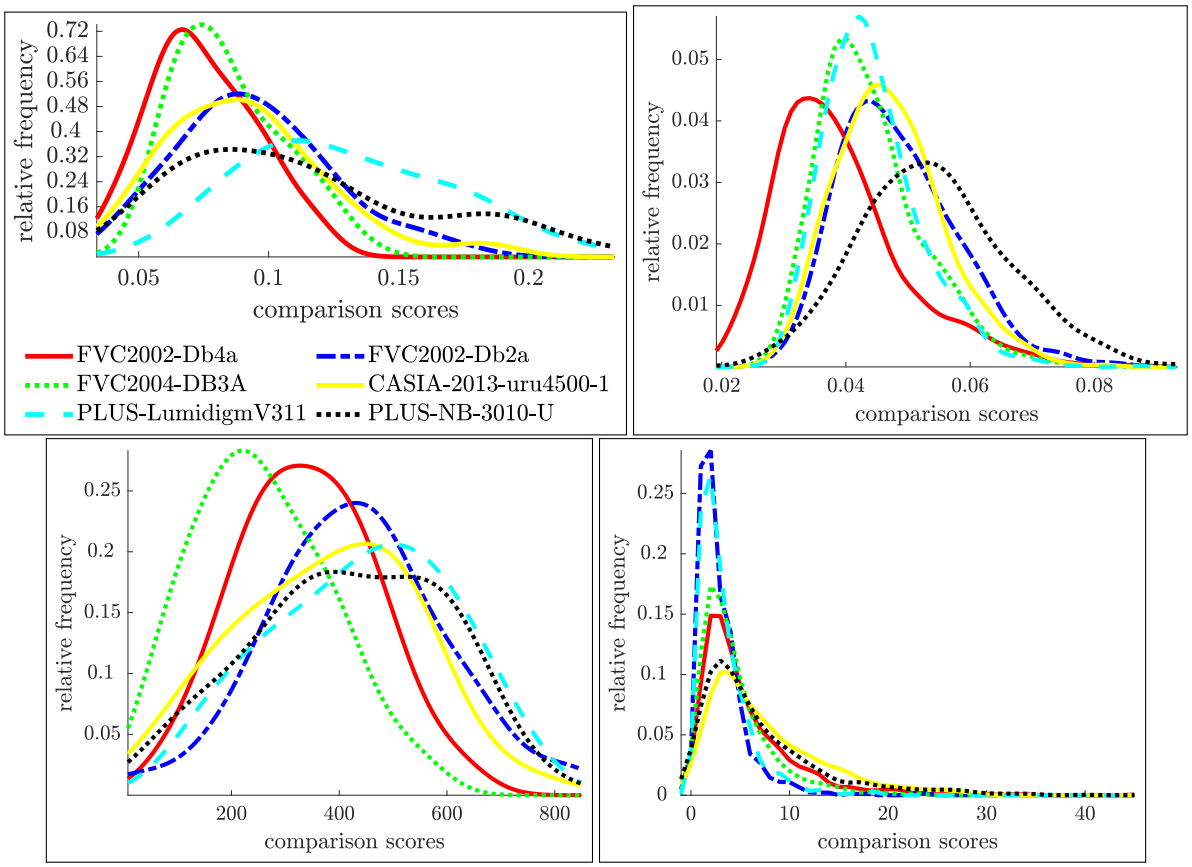

Figure 4: Comparison score distributions of samples from FVC, CASIA and PLUS datasets for genuine (left column) as well as imposter scores (right column) using FC (top row) and Verifinger (bottom row) as fingerprint recognition systems. The line colour, line style and naming of top left plot was used in all following distribution plots.

Table 4: P-values of Mann-Whitney-U tests of histogram metric values from real and synthetic score distribution comparisons. The null hypothesis in each case is defined such that the histogram metric values of real and synthetic score distribution comparisons are from continuous distributions with equal medians with an alpha significance level of 0.01.

	genuine	imposter	genuine	imposter
	FC		POC	
Chi	0.49	0.65	0.0015	0.11
HI	0.39	0.49	0.0013	0.13
KL	0.20	0.16	0.37	0.02
	ANSI		Verifinger	
Chi	0.47	0.69	0.77	0.0000
HI	0.46	0.94	0.95	0.0000
KL	0.22	0.59	0.41	0.0019

For all three metrics it is difficult so clearly separate the results for case a) and b). Thus, from a quality point-of-view it might be possible to differentiate between synthetic and real fingerprint samples in some cases by an in-depth analysis of the respective distributions by statistical means. In general, using one of the considered metrics does never lead to an unambiguous decision in terms of differentiating real vs. synthetic data, with the only exception shown in Table 3, where the p-values for a Mann-Whitney-U test [11] are presented. The Mann-Whitney-U test is a non-parametric test for two independent sample sets and allows a t-test identical interpretation of the results. However, it must be noted that the applied test is computed based on rank sums rather than means as it is done using a t-test. Here, the selected sample sets are on the one hand the metric values contained in the aforementioned case a) and on the other hand those contained in case b). If the values reported in Table 3 or Table 4 are lower than 0.01, the test indicates that the metric values have been drawn from different distributions, i.e. that synthetic and real samples can be differentiated using the respective metric. Only for one GSH case, which is shown in the left image of Figure 2, the test indicates that the KL values have been drawn from different distributions, i.e. that synthetic and real samples can be differentiated using the KL metric.

Overall, our results indicate that apart from the quality measure GSH when compared with the KL hsitogram metric, all settings indicate that the quality distributions of the synthetic data do not behave differently from the quality distributions of real data (as the

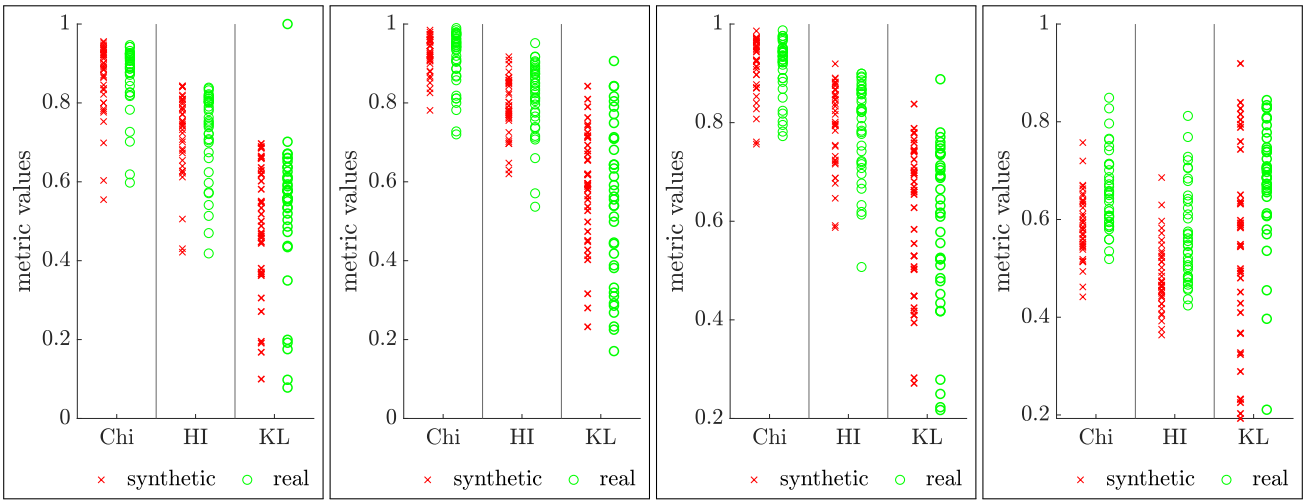

Figure 5: Plots showing the high overlap of histogram comparison metric results for genuine (left, center left plot) as well as imposter distributions (center right, right plot) using FC (left, center right plot) and Verifinger (center left, right plot) as fingerprint recognition systems.

latter also exhibit a significant variety in their respective characteristics). Therefore, we conclude that requirement R is fulfilled.

4.2.2 Diversity (D) and Controllable generation (C). Requirements D and C were evaluated using two minutiae and two non-minutiae based fingerprint recognition systems by generating genuine and imposter score distributions representing the comparison's behaviour of all four systems and performing technically the same analysis on the scores as done for requirement (R).

Score distributions are compared in Figure 4 where the first row corresponds to the averaged results using FC and the second one using Verifinger. In general, both minutiae-based systems, Verifinger and Innovatrics ANSI, exhibited a similar trend which only differs sightly from the non-minutiae based ones (MC and POC). There is a high overlap of imposter as well as genuine scores between both synthetic and real sample distributions.

Requirement D would be violated in case synthetic impostor score distributions would indicate a higher similarity among unrelated samples as compared to real data. This is clearly not the case in the examples shown. Thus, diversity (D) is given for the synthetic data considered. Contrasting to this, requirement C would be violated in case of strong overlap of genuine and impostor distributions, respectively. The FC score distributions exhibit a high overlap for both synthetic as well as real data – this indicates that FC scores are not suited to be used to assess requirement C. On the other hand, the Verifinger results do not show significant overlaps, neither for synthetic, nor for real data. Thus, controllable generation (C) is given for the synthetic data considered.

The histogram comparison metrics also report a high overlap of the distributions in general (see Figure 5). The only exceptions to this trend are the genuine distribution comparisons using FC (Chi and HI) and imposter scores using Verifinger evaluated by Chi and KL (the latter being relevant for assessing requirement D). In these cases some of the "synthetic vs real" sample comparison metric values can be clearly separated from the "real vs real" results. A statistical separation using the Mann-Whitney-U test (see Table 4), was not successful for the FC exceptions, while the Verifinger exceptions (impostor scores) are confirmed by the statistical analysis as well. According to the performed statistical test, Chi and HI for the genuine distributions evaluated by POC as well as all metrics for the impostor distributions evaluated by Verifinger lead to a separation of synthetic and real distributions. Hence, while a separation between synthetic and real fingerprint score distributions is not possible in general, it is feasible in some single cases, in particular for some specific recognition algorithms. However, while the impostor distributions are found to be different in these cases, they are not different in a way to indicate a violation of requirement D, see above for the corresponding discussion.

5 CONCLUSION AND FUTURE WORK

Our proposed requirements on synthetic fingerprint images (summarized in Table 2) are formulated towards establishing a consensus on how to evaluate the quality of a generative model that produces synthetic fingerprints or the quality of a dataset of synthetic fingerprints. Beyond the standard requirements for realistic appearance, sufficiently high image resolution, diversity and data anonymity, we ask for controllable generation, reflection of basic characteristics of ground truth (training) data for the case of data-driven generation and finally for equal distribution of privacy-related attributes such as gender, age, ethnicity, etc. in the generated samples. Moreover, for each requirement we propose a metric which indicates whether the requirement is met. We elaborate on which requirements are more important for which applications focusing on biometric authentication and forensic investigations. We draw a line between an informed evaluation of a generative model and a blind evaluation of an already generated dataset with respect to the proposed requirements.

In the experimental part, the proposed requirements are validated for several datasets of real and synthetic fingerprints to conclude whether it is possible to tell real and synthetic samples

(or datasets) apart using the proposed metrics. While doing so, we look into issues of realistic appearance, diversity and controllable generation.

Our future work will include research and development of GAN-based generative models capable of conditional generation of fingerprint images incl. embedding artifacts that characterize fingerprint sensor, substrate influence and environmental conditions as well as individual characteristics of a finger. We plan a full-stack evaluation of recently proposed fingerprint generators aiming at comparing them to each other and establishing their limits. Furthermore, we are going to study which consequences the violation of one or another requirement would have regarding privacy leakage as well as validity of the evaluation based on the synthetic data.

ACKNOWLEDGMENTS

The research in this paper is a part of the joint project GENSYNTH (Tools for the Generation of Synthetic Biometric Sample Data). The main contribution of the research group AMSL at the Otto von Guericke University Magdeburg is in Chapters 1 to 3, funded by the Deutsche Forschungsgemeinschaft (DFG) under project no. 421860227. The main contribution of the research group WaveLab at the University of Salzburg is in Chapter 4, funded by the Austrian Science Fund (FWF) under project no. I4272.

REFERENCES

[1] Sukesh Adiga V. and Jayanthi Sivaswamy. 2019. FPD-M-net: Fingerprint Image Denoising and Inpainting Using M-net Based Convolutional Neural Networks. In *Inpainting and Denoising Challenges*, Sergio Escalera, Stephane Ayache, Jun Wan, Meysam Madadi, Umut Güçlü, and Xavier Baró (Eds.). Springer Int. Publishing, Cham, 51–61.

[2] Mohamed Attia, MennattAllah H. Attia, Julie Iskander, Khaled Saleh, Darius Nahavandi, Ahmed Abobakr, Mohammed Hossny, and Saeid Nahavandi. 2019. Fingerprint Synthesis Via Latent Space Representation. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE Press, 1855–1861.

[3] Laurent Beslay and Javier Galbally. 2015. *Fingerprint identification technology for its implementation in the Schengen Information System II (SIS-II).* JRC Science for Policy Report. European Commision. Administrative Arrangement, JRC 33516-2014.

[4] Philip Bontrager, Aditi Roy, Julian Togelius, Nasir Memon, and Arun Ross. 2018. DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution. (2018). arXiv:cs.CV/1705.07386

[5] K. Cao and A. Jain. 2018. Fingerprint Synthesis: Evaluating Fingerprint Search at Scale. In *Proc. of the International Conference on Biometrics (ICB'18)*. 31–38.

[6] Raffaele Cappelli. 2009. SFinGe. In *Encyclopedia of Biometrics*, Stan Z. Li and Anil Jain (Eds.). Springer US, Boston, MA, 1169–1176.

[7] R. Cappelli, A. Lumini, D. Maio, and D. Maltoni. 2007. Fingerprint Image Reconstruction from Standard Templates. *IEEE Transactions on Pattern Analysis Machine Intelligence* 29 (2007), 1489–1503. Issue 9.

[8] Suhang Chen, Sheng Chang, Qijun Huang, Jin He, Hao Wang, and Qiangui Huang. 2014. SVM-Based Synthetic Fingerprint Discrimination Algorithm and Quantitative Optimization Strategy. *PLOS ONE* 9, 10 (2014), 1–9.

[9] Sergio Escalera, Martí Soler, Stephane Ayache, Umut Güçlü, Jun Wan, Meysam Madadi, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. 2019. ChaLearn Looking at People: Inpainting and Denoising Challenges. In *Inpainting and Denoising Challenges*, Sergio Escalera, Stephane Ayache, Jun Wan, Meysam Madadi, Umut Güçlü, and Xavier Baró (Eds.). Springer Int. Publishing, Cham, 23–44.

[10] Masud An-Nur Islam Fahim and H. Y. Jung. 2020. A Lightweight GAN Network for Large Scale Fingerprint Generation. *IEEE Access* 8 (2020), 92918–92928.

[11] Jean Dickinson Gibbons and Subhabrata Chakraborti. 2020. *Nonparametric statistical inference.* CRC press.

[12] Brigete Tatiana González. 2018. Analytical Comparison of Histogram Distance Measures. In *Proc. of the 23rd Iberoamerican Congress Progress in Pattern Recognition (CIARP'18)*, Vol. LNCS 11401. Springer.

[13] Carsten Gottschlich and Stephan Huckemann. 2014. Separating the real from the synthetic: minutiae histograms as fingerprints of fingerprints. *IET Biometrics* 3, 4 (2014), 291–301.

[14] J. Hämmerle-Uhl, M. Pober, and A. Uhl. 2013. Towards Standardised Fingerprint Matching Robustness Assessment: The StirMark Toolkit – Cross-Database Comparisons with Minutiae-based Matching. In *Proc. of the 1st ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'13)*. 111–116.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6629–6640.

[16] Mario Hildebrandt and Jana Dittmann. 2014. From StirMark to StirTrace: Benchmarking Pattern Recognition Based Printed Fingerprint Detection. In *Proc. of the 2nd ACM Workshop on Inf. Hiding and Multimedia Security (IH&MMSec'14)*. Association for Computing Machinery, New York, NY, USA, 71–76.

[17] Mario Hildebrandt and Jana Dittmann. 2015. StirTraceV2.0: Enhanced Benchmarking and Tuning of Printed Fingerprint Detection. *IEEE Transactions on Information Forensics and Security* 10 (2015), 833–848.

[18] Mario Hildebrandt and Jana Dittmann. 2016. StirTraceV3.0 and printed fingerprint detection: Simulation of acquisition condition tilting and its impact to latent fingerprint detection feature spaces for crime scene forgeries. In *Proc. of the 4th International Conference on Biometrics and Forensics (IWBF'16)*. 1–6.

[19] Anil K. Jain, Y. Chen, and M. Demirkus. 2006. Pores and Ridges: Fingerprint Matching Using Level 3 Features. In *Proc. of the 18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 4. 477–480.

[20] P. Johnson, F. Hua, and S. Schuckers. 2013. Texture Modeling for Synthetic Fingerprint Generation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 154–159.

[21] Ondrej Kanich. 2014. *Fingerprint damage simulation – a simulation of fingerprint distortion, damaged sensor, pressure, and moisture.* Lambert Academic Publishing.

[22] Ondřej Kanich, David Košťák, and Martin Drahanský. 2019. Psoriasis Damage Simulation into Synthetic Fingerprint. In *Proc. of the 18th Int. Conf. of the Biometrics Special Interest Group* (lecture notes in informatics ed.). GI - Group for computer science, 205–212.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proc. of Int. Conf. on Learning Representations (ICLR'18)*.

[24] Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. (2018). arXiv:1812.04948

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. (2019). arXiv:1912.04958

[26] Ronny Merkel and Mario Hildebrandtand Jana Dittmann. 2015. Application of StirTrace Benchmarking for The Evaluation of Latent Fingerprint Age Estimation Robustness. In *Proc. of the 3rd IEEE International Workshop on Biometrics and Forensics (IWBF'15)*. IEEE Press, Gjovik, Norway, 1–6.

[27] Shervin Minaee and Amirali Abdolrashidi. 2018. Finger-GAN: Generating Realistic Fingerprint Images Using Connectivity Imposed GAN. (2018). arXiv:cs.CV/1812.10482

[28] Vishesh Mistry, Joshua J. Engelsma, and Anil K. Jain. 2020. Fingerprint Synthesis: Search with 100 Million Prints. (2020). arXiv:cs.CV/1912.07195

[29] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit B. Patel, and Anima Anandkumar. 2020. Semi-Supervised StyleGAN for Disentanglement Learning. (2020). arXiv:cs.CV/2003.03461

[30] U.S. Department of Justice. 2012. *The Fingerprint Sourcebook.* CreateSpace Independent Publishing Platform.

[31] U.S. National Institute of Standards and Technology. 2016. *NFIQ2.0: NIST Fingerprint Image Quality, Interagency Report XXXX.* Draft.

[32] Martin Aastrup Olsen, Vladimír Šmida, and Christoph Busch. 2016. Finger image quality assessment features–definitions and evaluation. *IET Biometrics* 5, 2 (2016), 47–64.

[33] M. Sadegh Riazi, Seyed M. Chavoshian, and Farinaz Koushanfar. 2020. SynFi: Automatic Synthetic Fingerprint Generation. (2020). arXiv:eess.IV/2002.08900

[34] Stefan Seidlitz., Kris Jürgens., Andrey Makrushin., Christian Kraetzer., and Jana Dittmann. 2021. Generation of Privacy-friendly Datasets of Latent Fingerprint Images using Generative Adversarial Networks. In *Proc. of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*. INSTICC, SciTePress, 345–352.

[35] Elham Tabassi, Charles Wilson, and Craig I. Watson. 2004. *NIST Fingerprint Image Quality.* Technical Report. NISTIR 7151.

[36] Aleksei Triastcyn and Boi Faltings. 2019. Generating Artificial Data for Private Deep Learning. (2019). arXiv:cs.LG/1803.03148

[37] Nicholas Whiskerd, J. Dittmann, and C. Vielhauer. 2018. A Requirement Analysis for Privacy Preserving Biometrics in View of Universal Human Rights and Data Protection Regulation. In *Proc. of the 26th European Signal Processing Conference (EUSIPCO'18)*. 548–552.

[38] André Brasil Vieira Wyzykowski, Mauricio Pamplona Segundo, and Rubisley de Paula Lemes. 2020. Level Three Synthetic Fingerprint Generation. (2020). arXiv:cs.CV/2002.03809