Time flies by: Analyzing the Impact of Face Ageing on the Recognition Performance with Synthetic Data

1st Marcel Grimmer
Norwegian Biometrics Laboratory
NTNU
Gjøvik, Norway
marceg@ntnu.no

2nd Haoyu Zhang
Norwegian Biometrics Laboratory
NTNU
Gjøvik, Norway
haoyu.zhang@ntnu.no

3rd Raghavendra Ramachandra
Norwegian Biometrics Laboratory
NTNU
Gjøvik, Norway
raghavendra.ramachandra@ntnu.no

4th Kiran Raja
Norwegian Biometrics Laboratory
NTNU
Gjøvik, Norway
kiran.raja

5th Christoph Busch
Biometric and Internet-Security Research Group
Hochschule Darmstadt
Darmstadt, Germany
christoph.busch@h-da.de

Abstract—The vast progress in synthetic image synthesis enables the generation of facial images in high resolution and photorealism. In biometric applications, the main motivation for using synthetic data is to solve the shortage of publicly-available biometric data while reducing privacy risks when processing such sensitive information. These advantages are exploited in this work by simulating human face ageing with recent face age modification algorithms to generate mated samples, thereby studying the impact of ageing on the performance of an open-source biometric recognition system. Further, a real dataset is used to evaluate the effects of short-term ageing, comparing the biometric performance to the synthetic domain. The main findings indicate that short-term ageing in the range of 1-5 years has only minor effects on the general recognition performance. However, the correct verification of mated faces with long-term age differences beyond 20 years poses still a significant challenge and requires further investigation.

Index Terms—Synthetic Data, Face Age Modification, Face Recognition

I. INTRODUCTION

The deployment of face recognition systems has gained popularity in various application scenarios, such as border control initiatives like the European Entry-Exit System (EES) [15]. In particular, the EES will be used as a central system for collecting and querying traveller data to the Schengen area at all border crossing points to facilitate the cooperation of visa and law enforcement authorities. The biometric performance of a system deployed in such sensitive environments must comply with high standards, such as those defined in the best practices for automated border control of the European Border and Coast Guard Agency (Frontex) [7]. At the same time, the European General Data Protection Law complicates the processing of biometric data to avoid privacy leakages.

Without an appropriate performance testing strategy, the risk of security lapses increases significantly and allows for the discriminatory treatment of travellers due to algorithmic or dataset bias. One solution to the lack of available test data includes the generation of synthetic data samples. However, in order to conduct reliable biometric performance tests, the synthetic samples must be as similar as possible to data collected in operational environments.

In the context of synthetic face images, the main focus of this work is to analyse the impact of human face ageing on biometric recognition performance. Due to the 10-year validity of EU passports and enrolment records in immigration systems, face recognition engines employed at the EU borders are frequently exposed to mated face comparisons captured over long time spans. This work deepens the understanding of recognition accuracy and face ageing by analysing synthetically generated face images rendered with ageing effects. This work relies on face age modification methods to avoid the time-consuming data collection of mated samples over time.

This analysis is based on face age manipulation frameworks operating within the latent space of StyleGAN [9] and StyleGAN2 [2]: InterFaceGAN [11] and SAM [17]. The choice of these techniques is motivated by the high realism and resolution (1024x1024) of facial images the StyleGAN generator achieves. The age-modified face images are analysed with two different face quality assessment algorithms (FQAs): FaceQnet v1 [3] and SER-FIQ [8]. The biometric performance is further evaluated by computing mated and non-mated comparison scores with ArcFace [4]. The breakdown of mated comparison scores into age bins enables precise testing of the weaknesses of existing face recognition engines.
Further, the UNCW face ageing dataset (also MORPH-II) [6] is used as a reference for comparing short-term ageing effects to those ageing effects achieved in the synthetic domain. This work is structured as follows: A brief introduction of the face age modification frameworks used to generate the synthetic datasets is given in Section II. The characteristics of the synthetic and reference datasets are described in Section III. Finally, the experimental results are presented in Section IV, analysing the biometric quality and the comparison scores based on synthetic and bona fide data.

II. FACE AGE MODIFICATION

This section introduces the basic terms and methods used to create the synthetic cross-age datasets analysed in this work. Face age progression (FAP) refers to rendering from a given input image a synthetic face image with ageing effects, while face age regression (FAR) corresponds to the prediction of rejuvenation effects [16]. Typically, recent face age modification (FAM) methods predict the appearance of an individual based on a given target age. Another type of FAM technique focuses on changing the age of subjects on a continuous scale with the motivation to better approximate the nature of human ageing. This work evaluates the impact of face ageing on a face recognition (FR) system, using two state-of-the-art FAM frameworks: SAM [17] and InterFaceGAN [11]. Both FAM frameworks are based on manipulating latent vectors in the latent space of StyleGAN [9] and StyleGAN2 [2]. The main idea is to exploit the disentanglement of facial attributes given in the internal data representation of a generative adversarial network (GAN). Operating directly in the latent space of a pre-trained GAN alleviates the need to train complex adversarial networks and benefits from the high resolution and photorealism achieved by the StyleGAN generators.

The basic FAM principles of SAM and InterFaceGAN are illustrated in Figure 1. The main question is where to move the randomly drawn latent vector to change the age while leaving other facial attributes unchanged [1]. InterFaceGAN addresses this issue by training a binary age boundary that divides the latent space into two subspaces (old vs young). Afterwards, the age is increased by moving an arbitrary latent vector into the perpendicular direction of the age boundary, with the magnitude defining the ageing extent. Unlike InterFaceGAN, SAM trains an additional age encoder conditioned on the target age \( \alpha_t \), extracting the missing ageing patterns by learning the residuals to the original face image. In a next step, a pre-trained map2style network [5] transforms the residual ageing patterns into latent codes, which are then fused with the initial latent vector randomly drawn from the latent space. After fusing the residual age patterns with the initial latent vector, the resulting latent code is passed to the StyleGAN2 generator to generate the age-modified face image.

III. DATASETS

This section introduces the synthetic cross-age datasets generated with InterFaceGAN and SAM, as well as the bona fide reference datasets (FRGC v2.0 [12], UNCW ageing dataset [6]).

A. Synthetic Dataset Generation

Our base synthetic data is randomly generated by the StyleGAN [9] generator pre-trained on the FFHQ [9] dataset. Choosing a truncation factor of \( \psi = 0.75 \) has proven as an effective setup [10] for generating visually appealing face images with a high diversity of demographic factors. Further, the work of Zhang et al. [10] indicates minor differences in the face recognition performance between StyleGAN and StyleGAN2 generated face images. Therefore, we select a dataset of 50,000 face images generated with StyleGAN as a basis for our face age modification algorithms. Given these synthetic base images, the corresponding age-modified samples are generated using the proposed semantic editing algorithms of Shen et al. [11] (InterFaceGAN) and Alaluf et al. [17]. As shown in Figure 1, InterFaceGAN controls the shifting distance in the latent space with a scaling factor that we empirically set as \( s_1 \pm 0.4, s_2 \pm 0.8, s_3 \pm 1.2 \) to create 6 synthetic data subsets containing mated samples of the base images.

As introduced in Section II, the input of SAM [17] is a target age and a base image. We select 7 different target age groups (10, 20, 30, 40, 50, 60, 70) to which we transform the base synthetic images to. In the original SAM algorithm, the pix2style2pixel (pSp) [5] encoder is applied to first project the base images into the extended StyleGAN2 latent space.
(W+) in order to fuse it with the encoded age residual code. In this work, we discard the initial base image as soon as it is projected to the W+ latent space, re-defining the reconstructed face image as our new base in order to avoid the distortion of our results due to identity losses caused by GAN inversion. In this work, face images with unrealistic capturing conditions are filtered out to increase the representativeness of our datasets. Details of the filtering pipeline are given in Table I. For the inter-eye-distance (IED), a pre-trained landmark detection model is used to predict the centre of the eyes and filter out images with IED less than 90 pixels or failed landmark detections. To filter out images with unsatisfying illumination conditions, we prepare an internal dataset with binary labels (good or poor illumination condition) and train a random forest regressor on the extracted features that measure illumination uniformity and symmetry from these images. The Img2pose model [13] is applied to predict the Euler angles of the head pose and filter out images with extreme rotations. Additionally, we included C3AE [18] to predict the age of the given images and exclude those with ages not in the range of [13,59] years. Finally, Table I illustrates the exact number of face images in the analysed datasets - before and after applying the filtering pipeline.

### B. Bona Fide Reference Datasets

To compare the synthetic data with real data, we choose a representative dataset containing 17,919 images from FRGC-V2 [12], which is known for its high-quality images and constrained conditions resembling those of border crossing capturing environments. However, despite the good representativeness, FRGC-V2 samples are not annotated with ground-truth ages, thus limiting the age-based performance comparison to the synthetic datasets. To overcome this limitation, the UNCW face ageing dataset [6] is further used in our analysis, including more than 55,000 face images of more than 13,000 individuals with exact age annotations, where for mated comparison trials the difference in age is ranging from 164 days to 1,681 days. In order to analyse short-term ageing effects and their impact on the face recognition performance, we sorted out mated samples with less than 1 year passed between the probe and reference image capturing, leaving an amount of 37,423 face images. Similarly, the synthetic datasets have been further reduced to only include mated pairs with age differences within 1 to 5 years. As no age labels are given for the synthetic data, we apply the C3AE [18] age estimator to predict the age labels for each face image individually. The resulting short-term InterFaceGAN (ST-InterFaceGAN) and SAM (ST-SAM) datasets comprise 10,772 and 12,298 samples.

### IV. EXPERIMENTAL RESULTS

#### A. Face Image Quality Assessment

The main goal of this subsection is to compare the age-modified datasets with the reference bona fide datasets by utilising face quality assessment algorithms (FQAAs). FQAAs are developed to predict the biometric quality of a given face image by translating its suitability for face recognition to a scalar value between [0, 1] (1= “Perfect biometric face image quality”, 0=“worst biometric quality”). In particular, two well-established deep learning-based FQAAs are selected for this task: FaceQnet v1 [3] and SER-FIQ [8]. In Figure 2, the notched boxplots visualise the median face image qualities with their 95%-confidence intervals. In addition, the horizontal red line represents the median of the FRGC-V2 reference dataset. This view enables the analysis of statistical deviations of the synthetic datasets’ medians to the median biometric quality of real data. In this context, Figure 2 reveals that all boxplots enclose the red line, thus indicating no statistical differences in the biometric quality across all age-modified datasets. However, it is noticeable that the medians of the synthetic datasets estimated with SER-FIQ consistently falls below the red line, thus supporting the conclusion of minor biometric quality differences between synthetic and real data. In contrast, the medians estimated with FaceQnet v1 fluctuate below and above the red reference line, hence strengthening the ”no difference” hypothesis.

#### B. Comparison Score Analysis

This section analyses the mated and non-mated comparison scores obtained by comparing each base image to the age-modified versions (mated) and other age-modified identities (non-mated). Figure 3 shows that increasing age modifications lead to decreasing similarities between the synthetic mated samples generated with InterFaceGAN. This observation corresponds to our initial hypothesis that existing face recognition systems are not trained to compensate ageing effects. The more years pass between the enrolment process and the recapturing of a probe image, the more intense ageing effects will occur and thus affect the recognition performance. On the contrary, the non-mated comparison scores illustrate only minor performance differences, as seen by the nearly identical distributions.

Similarly, Figure 3 shows the kernel density plots of comparison scores obtained by comparing synthetic samples generated with SAM. The general behaviour of the mated and non-mated comparison scores is similar to the results reported with InterFaceGAN. The more the target age differs from the average StyleGAN age (34y), the less similar the mated samples become. This loss in identity over time can either be caused by the FAM algorithms or the incapability of ArcFace

| Dataset          | Images before Filtering | Images after Filtering |
|------------------|-------------------------|------------------------|
| FRGC v2          | 23,705                  | 17,919                 |
| Synthetic Base   | 48,428                  | 23,705                 |
| InterFaceGAN (scale = 0.4) | 18,186                | 48,428                 |
| InterFaceGAN (scale = 0.8) | 18,186                | 48,428                 |
| InterFaceGAN (scale = 1.2) | 18,186                | 48,428                 |
| VMM (target age = 10) | 25,918                 | 18,290                 |
| VMM (target age = 20) | 25,918                 | 22,671                 |
| VMM (target age = 30) | 25,918                 | 23,253                 |
| VMM (target age = 40) | 25,918                 | 23,253                 |
| VMM (target age = 50) | 25,918                 | 22,671                 |
| VMM (target age = 60) | 25,918                 | 17,174                 |
| VMM (target age = 70) | 25,918                 | 10,528                 |

**TABLE I: General Database Information**
Fig. 2: Biometric quality analysis of synthetic images generated with InterFaceGAN (top) and SAM (bottom). The biometric quality is estimated with two FQAAs: SER-FIQ (right) and FaceQnet v1 (left). The red line visualises the median biometric quality of the bona fide reference dataset (FRGC-V2).

Fig. 3: Mated (left column) and non-mated (right column) comparison scores based on age-modified datasets generated with InterFaceGAN (top) and SAM (bottom)
to handle long-term age differences between the probe and reference sample. That to say, the disentanglement of these two sources of potential identity loss remains a challenging task.

1) Synthetic vs Natural Face Ageing: The final part of this section analyses the impact of short-term ageing effects based on the ST-ageing datasets introduced in Section III-B. Figure 4 shows the mated comparison score distributions, indicating no significant differences between mated samples collected without ageing (FRGC v2.0) and those with age gaps within 1 to 5 years (UNCW). The main reason for this observation is most likely due to minor ageing patterns given the short time intervals available in this analysis. A similar behaviour is observed with SAM generated face images (red line), hence indicating that the synthetic mated samples are similar to those seen in real data.

Finally, the cyan curve achieved the highest similarity scores, emphasising the capability of InterFaceGAN to preserve identity information during the age synthesis. Despite the effective identity preservation rate, the comparison to the real curves reveals a large domain gap, thus potentially overestimating mated comparison scores observed in real-world settings. Another domain-gap crystallizes in the non-mated comparison scores in Figure 4: While the synthetic lines are nearly identical, their average scores are higher than those measured for bona fide data.

V. CONCLUSION

The main focus of this work is to analyse the impact of face ageing on face recognition systems by using FAM to generate synthetic mated and non-mated samples with varying age gaps. For this purpose, the FAM performance is analysed in terms of the biometric quality (Section IV-A) and identity preservation (Section IV-B) of the generated mated and non-mated face images. The main findings of this work underline the capability of synthetic face images to interfere with face recognition systems similar than bona fide data. Further, the comparison score analysis indicates only a minor deterioration in the recognition performance for short and medium-term ageing intervals - as shown by comparisons conducted in the synthetic and real domain. Nevertheless, the mated comparison scores significantly decrease for long-term age intervals or extreme target age choices. Finally, this work demonstrates the future value of synthetic face images in analysing the age-robustness of FR systems. Accelerated by the remarkable progress of deep generative networks, we believe the domain gap between synthetic and bona fide data to vanish over time. Especially in the context of face ageing, FAM algorithms are crucial for avoiding long-lasting data collection initiatives, which are not feasible given the time constraints of real-world applications. In an endeavour of closing the domain gap between synthetic and bona fide data, future research may benefit from new concepts, such as 3D multi-view image synthesis [14] in order to better preserve spatial information of the faces and support the generation of geometry-consistent mated samples.

REFERENCES

[1] M. Grimmer, H. Zhang, R. Ramachandra, K. Raja and C. Busch “Generation of Non-Deterministic Synthetic Face Datasets Guided by Identity Priors,” 33rd Norwegian Information Security Conference, 2021.
[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila “Analyzing and improving the image quality of stylegan,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020.
[3] J. Hernandez-Ortega, J. Galbally, J. Fierrez and L. Beslay “Biometric quality: Review and application to face recognition with faceqnet,” arXiv preprint arXiv:2006.03298, 2020.
[4] J. Deng, J. Guo, N. Xue and S. Zafeiriou “Arcface: Additive angular margin loss for deep face recognition,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019.
[5] E. Richardson Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro and D. Cohen-Or “Encoding in style: a stylegan encoder for image-to-image translation,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2021.
[6] K. Rianek and T. Tesafaye “Morph: A longitudinal image database of normal adult age-progression,” 7th Intl. Conf. on Automatic Face and Gesture Recognition, 2006.
[7] Frontex “Best practice technical guidelines for Automated Border Control (ABC) systems,” European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, 2015.
[8] P. Terhorst, J.N. Kolf, N. Damer, F. Kirchbuchner and A. Kuiper “SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness,” Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[9] T. Karras, S. Laine and T. Aila “A style-based generator architecture for generative adversarial networks,” Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[10] H. Zhang M. Grimmer, R. Ramachandra Raja, Kiran and C. Busch “On the Applicability of Synthetic Data for Face Recognition,” IEEE Intl. Workshop on Biometrics and Forensics, 2021.

[11] Y. Shen, C. Yang, X. Tang and B. Zhou “Interfacegan: Interpreting the disentangled face representation learned by gans,” IEEE transactions on pattern analysis and machine intelligence, 2020.

[12] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek “Overview of the face recognition grand challenge,” IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2005.

[13] V. Albiero, X. Chen, X. Yin, G. Pang and T. Hassner “img2pose: Face alignment and detection via 6dof, face pose estimation,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2021.

[14] E.R. Chan, C.Z. Lin, M.A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L.J. Guibas, J. Tremblay, S. Khamis and others “Efficient geometry-aware 3D generative adversarial networks,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2022.

[15] European Council “Commission Implementing Decision 2019/329 of 25 February 2019 laying down the specifications for the quality, resolution and use of fingerprints and facial image for biometric verification and identification in the Entry/Exit System (EES),” 2019.

[16] M. Grimmer, R. Ramachandra and C. Busch “Deep face age progression: A survey,” IEEE Access, 2021.

[17] Y. Alaluf, O. Patashnik and D. Cohen-Or “Only a Matter of Style: Age Transformation Using a Style-Based Regression Model,” ACM Trans. Graph., 2021.

[18] M. Grimmer, R. Ramachandra and C. Busch “C3AE: Exploring the limits of compact model for age estimation,” Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019.