Validating Seed Data Samples for Synthetic Identities – Methodology and Uniqueness Metrics

VIKTOR VARKARAKIS, (Graduate Student Member, IEEE), SHABAB BAZRAFKAN, (Student Member, IEEE), GABRIEL COSTACHE, AND PETER CORCORAN, (Fellow, IEEE)

1Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
2Imec Vision Laboratory, Department of Physics, University of Antwerp, 2000 Antwerp, Belgium
3Xperi, Galway, H91 V0TX Ireland

Corresponding author: Viktor Varkarakis (v.varkarakis1@nuigalway.ie)

This work was supported in part by the Science Foundation Ireland (SFI) through the SFI Strategic Partnership Program, and in part by FotoNation Ltd., on Next-Generation Imaging for Smartphone and Embedded Platforms, under Project ID:13/SPP/12868.

ABSTRACT This work explores the identity attribute of synthetic face samples derived from Generative Adversarial Networks. The goal is to determine if individual samples are unique in terms of identity, firstly with respect to the seed dataset that trains the GAN model and secondly with respect to other synthetic face samples. Two approaches are introduced to enable the comparative analysis of large sets of synthetic face samples. The first of these uses ROC curves to determine identity uniqueness using a number of large publicly available datasets of real facial samples to provide reference ROCs as a baseline. The second approach uses a thresholding technique utilizing again large publicly available datasets as a reference. For this approach, new metrics are introduced, and a technique is provided to remove the most connected data samples within a large synthetic dataset. The remaining synthetic samples can be considered as unique as data samples gathered from different real individuals. Several StyleGAN models are used to create the synthetic datasets, and variations in key model parameters are explored. It is concluded that the resulting synthetic data samples exhibit excellent uniqueness when compared with the original training dataset, but significantly less uniqueness when comparisons are made within the synthetic dataset. Nevertheless, it is possible to remove the most highly connected synthetic data samples. Thus, in some cases, up to 92% of the data samples in a 20k synthetic dataset can be shown to exhibit similar uniqueness to data samples taken from real public datasets.

INDEX TERMS Artificial intelligence, computer vision, face recognition, generative adversarial networks (GANs), StyleGAN, synthetic face, synthetic identity, uniqueness metrics.

I. INTRODUCTION

In the last few years, a number of tools for generating synthetic facial samples have evolved [1]–[3], based on generative adversarial networks (GANs) [4]. These enable photo-realistic, high-resolution synthetic face samples to be generated at scale. StyleGAN [3] is a representative of the current state-of-art, and the generated samples are photo-realistic and of higher quality than the facial samples available in many public face datasets. This leads us to consider the potential to create a large facial dataset built entirely from synthetic facial data samples.

Now a key attribute of face samples is their association with a specific person or individual. We refer to this association as the identity of a face sample. With the introduction of GDPR in Europe and similar data privacy regulations in other jurisdictions, it has become challenging to gather biometric data, in particular, facial data for research purposes. Facial data samples are directly associated with an individual, and it is challenging to anonymize or otherwise separate facial data from a person’s identity. Consequently, some biometric datasets have been withdrawn from public use, and building a new dataset has become increasingly complex and expensive.

A. MOTIVATION & RESEARCH QUESTIONS

This work explores the potential to build synthetic facial datasets at scale by using a GAN to generate a seed dataset of facial data samples that are demonstrably unique in terms of their identity. Given such a seed dataset, it would then be
feasible to modify these seed samples to build large synthetic training datasets focusing on other facial attributes such as facial lighting, pose, and expression [5]–[7].

The starting point for building such datasets is a methodology to demonstrate that the identities of the synthetic facial data samples used as seed data behave in the same way as those of a ‘real-world’ dataset of facial data samples. It is also essential to validate that the synthetic data samples are unique in terms of identity with the original seed data used to train the generator. These considerations lead to three key research questions:

1) Are the synthetic data samples unique when compared with the original seed data used to train the GAN model?

2) Are the synthetic data samples within a generated dataset unique when compared with one another?

3) Can we validate individual samples within a generated dataset to ensure that there is sufficient identity uniqueness to use as a synthetic seed data sample for further research?

B. APPROACH AND METHODOLOGY

These research questions led us to develop two research approaches to understand and quantify the identity uniqueness within a set of samples of synthetic (Sy) data when compared against samples from the seed (Se) dataset (Sy vs Se). The same methodology can then be applied to understand the uniqueness within a set of synthetic data samples, in terms of identity (Sy vs Sy). To evaluate the identity uniqueness of the generated synthetic samples for both cases (Sy vs Se, Sy vs Sy), the proposed approaches utilize a state-of-art Face Recognition (FR) model. In both approaches, the performance/behavior of real samples is used as a reference point, and the performance/behavior of the generated synthetic samples for each case is compared against it to draw conclusions and answer the Research Questions.

In the first approach, the performance/behavior of real samples and generated synthetic samples for each examined case (Sy vs Se, Sy vs Sy) is illustrated through ROC curves. These are compared and examine the identity uniqueness of the generated synthetic samples with their seed data and identity uniqueness of the generated synthetic samples when compared with one another.

In the second approach, a thresholding technique is implemented to determine the identity uniqueness for both cases. In this approach, the performance of real samples is illustrated through an FR threshold. The FR threshold is used to determine the similarity of the generated samples with their seed dataset and also among the synthetic samples in terms of identity. Reversely, this shows the identity uniqueness. Using this approach, a new metric is introduced which based on its value, answers the questions posed on this work along with an approach to quantify the generated synthetic samples that have a unique identity in each case (Sy vs Se, Sy vs Sy).

The paper is structured in the following way. A Literature Review is initially given along with the Foundation Methods used in this work. The Methodology is described, followed by its Implementation and Experiments. Finally, the results are discussed in the Conclusion, along with Future Work.

II. LITERATURE REVIEW

Several evaluation measures have surfaced with the emergence of new GAN models. Some of them attempt to quantitatively evaluate models while others emphasize on qualitative ways such as studies or analyzing internals of models [8].

Regarding quantitative metrics, the Inception Score (IS) proposed in [9], is one of the most popular scores for evaluating GAN models [10]. In order to compute the IS, the generated images are passed through Inception Net [11] (trained on ImageNet [12]) and the output is post-processed to capture different properties of the image. The IS score is able to show a reasonable correlation with the quality and diversity of generated images [11]. Other metrics have also been introduced, which use similar concepts as in IS, such as M-IS [13], Mode Score [14], AM score [15], and FID [16].

Also, a common indirect technique for GAN evaluation, especially for Conditional-GANs, is to use an off-the-shelf classifier to assess the synthetic images [8]. For example, in [17], a VGG network was utilized to evaluate the fake colored images. This method is called semantic interpretability. A similar approach was used in [18], where the FCN score is proposed to measure the quality of the generated images and also in [19], where the GAN Quality Index (GQI) is introduced. Finally, researchers have proposed measures from the image quality assessment literature, such as SSIM, PSNR, or/and Sharpness Difference (SD), to be used not only in evaluating the GAN models but also in training [18], [20]–[22].

The qualitative metrics used to evaluate GAN models are divided into 5 main categories in [8]: (i) Nearest Neighbors approaches to detect overfitting [20], [23], (ii) Rapid Scene Categorization methods, in which humans are reporting features of the generated images with a quick look [24]–[26], (iii) Rating and Preference Judgment, where humans rate the synthetic images in terms of fidelity [17], [21], [27]–[32], (iv) approaches where the mode drop/collapse of the GANs models are examined [33]–[35] and finally (v) methods of Investigating and Visualizing the Internals of Networks, to explore what and how the GAN models learn through latent space exploration [1], [36]–[41].

All approaches have strengths and limitations, which are discussed extensively in [8]. Even the IS and FID, have drawbacks as they rely on pre-trained deep networks to represent and statistically compare original and generated samples and using a certain natural scene dataset (e.g., ImageNet), and applying them to other domains is questionable [8]. However, these two metrics are widely accepted in evaluating GAN models. These, along with other issues, have made evaluating generative models notoriously difficult [23] and there exists no agreement regarding the best GAN evaluation measure [8].

Due to these challenges, it is argued against evaluating models for task-independent image generation and proposed to evaluate GANs with respect to a specific application as for different applications different measures might be
appropriate [23]. This work’s Methodology focuses on the uniqueness of the identity attribute of the synthetic face samples. Therefore, it is only applicable to GAN models trained for the task of face generation. The proposed Methodology enables us to understand if the generate synthetic face samples are unique, in terms of identity, when compared to their seed data samples. The approach is also applied to examine the uniqueness among the generated synthetic data sample. As an extension, this Methodology allows us to quantify the synthetic data with a unique identity when compared to their seed data or with other synthetic samples, which can be used to measure the ability of a GAN to generate synthetic data with unique identities.

III. FOUNDATION TECHNIQUES
This section presents the Foundation Techniques employed in this research, including an introduction to the GAN model selected in generating the synthetic samples that are examined, the datasets used, the employed face recognition model as well as the measurement techniques used to generate the proposed Methodology.

A. GENERATING SYNTHETIC FACIAL DATA WITH GANS
Currently, StyleGAN [3] represents the state-of-the-art GAN for the face generation task and the synthetic facial samples that are examined in this work are derived from it. Although any GAN model trained on the task of face generation can be used to implement the Methodology proposed. Three different StyleGAN models are used for the result of this study to not rely on a single model. The datasets that each model has been trained on has a different number of samples and a different number of identities. Note that in this work, we did not consider other GANs due to the complexity and computation effort of the re-training process, but it would be interesting to investigate and compare other GANs that are used for facial generation and this is commented on in the future work section.

When considering the distribution of the training data, areas of low density are poorly represented and thus likely to be difficult for the generator to learn which is a significant open problem in all generative modeling techniques [3]. However, it is known that drawing latent vectors from a truncated [42], [43], or otherwise shrunk [44] sampling space tends to improve average image quality, although some amount of variation is lost. To avoid generating poor images, StyleGAN [3] truncates the intermediate vector w, forcing it to stay close to the “average” intermediate vector. The truncation psi value ranges from {−1, 1} and influences how diverse the output will be. The further the truncation psi value is from 0, the less truncated (more diverse) the sampling space is. In the work presented, the influence of StyleGAN’s truncation psi parameter is studied, and sets of synthetic data samples are created using different truncation psi values. More details regarding the architecture, as well as the hyperparameter selection of StyleGAN, can be found in [3].

In the Implementation and Experiment, section V, the StyleGAN models used in this work are described. Also, the procedure and details used to generate the synthetic data are given.

B. DATASETS
The following datasets are used as part of this research for training and evaluation purposes.

1) LABELED FACES IN THE WILD (LFW)
Labeled Faces in the Wild (LFW) [45] is the de facto standard test dataset for the face verification in unconstrained conditions. The majority of research publications related to the face verification task report their performance with the mean face verification accuracy and the ROC curve on the standard evaluation set of 6,000 given face pairs in LFW. The dataset was released in 2007 and contains 13,233 face images of 5,749 identities. Although, it should be mentioned that due to the small number of identities and the number of samples per identity in the LFW, it is inadequate for training purposes and thus is used mainly for testing. In this work, the LFW is utilized to compute ROC curves which are part of the proposed Methodology. In the Implementation and Experiment, section V, it is explained in detail how it is used.

2) CASIA-WEBFACE
CASIA-WebFace [46] is one of the first large public facial datasets, published in 2014. It contains 10,575 identities, with a total of 494,414 facial data samples. The identities belong to celebrities and all of them are collected from the IMDb website. The size of the dataset makes it suitable for facial recognition tasks and this dataset is frequently used as a baseline by researchers in the facial recognition field. CASIA-WebFace is used similarly as the LFW in this work and more details can be found in the Implementation and Experiment section V.

3) CELEBFACES/CELEBA
The CelebFaces+ dataset was released in 2014 and along with the CASIA-WebFace was one of the first large publicly available datasets, as it contains 202,599 images of 10,177 identities. A version of this dataset with additional metadata is known as CelebFaces Attributes Dataset (CelebA) [47], where the samples form CelebFaces+ are annotated with 5 landmark locations and for 40 binary attributes (eyeglasses, mustache, hat, etc.), providing valuable information for the researchers. CelebA has a two-fold use in this work. As LFW and CASIA-WebFace, it is used similarly in order for ROC curves to be computed. In addition, it is used in training a StyleGAN model from which synthetic data are generated and examined in this work.

4) CELEBA-HQ
CelebA-HQ [48] is a high-quality subset version of the CelebA dataset. Consists of 30,000 face samples in 1024 × 1024 resolution. The original samples from CelebA
were utilized and pre-processed, in order to achieve consistent high quality and center the images on the facial region. The pre-processing pipeline used to produce the CelebA-HQ from the CelebA dataset is described in [48]. The dataset was created and used initially to train PGAN [48] and also StyleGAN [3]. The 30,000 samples are from approximately 6,000 identities. The StyleGAN model trained on the CelebA-HQ from [3], is used to generate synthetic samples which are used to answer the Research Questions posed in the Introduction.

5) FFHQ
Flickr-Faces-HQ (FFHQ) [3] is a high-quality image dataset of human faces. The datasets consist of 70,000 high-quality images at a resolution of 1024 × 1024. The samples were crawled from Flickr. The dataset has considerable variation in terms of ethnicity, age, image background and accessories. The dataset is created as a benchmark for generative adversarial networks and each face sample originates from a different person. The FFHQ is used to train a StyleGAN model [3], from which synthetic samples are created which are used in the Methodology of this work.

C. FACE RECOGNITION MODEL
The face recognition (FR) model selected for use in this work is the ArcFace [49]. ArcFace was made public in 2018 and the results presented at that time pushed the limits of the LFW benchmark beyond state-of-art at that time, achieving 99.83% accuracy. It also achieved state-of-the-art results on the MegaFace Challenge [50].

The proposed methodology can be implemented using any FR model with the condition that should be a state-of-art model and having a high performance on the datasets used to implement the Methodology. The choice of the ArcFace model is because of the availability, as the authors have released the weights of the model. Other state-of-art FR models such as FaceNet [51] or CosFace [52] do not provide official implementations with reference training weights. Before concluding in the use of the ArcFace model, an unofficial implementation of FaceNet was also tested but it didn’t have a high performance on the datasets used in this work. Therefore extra fine-tuning should have been implemented and possible making the FR model biased on a specific dataset. On the contrary, the ArcFace model has high performance on the datasets used in this work without any fine-tuning or initial training on them. Thus, the use of ArcFace will enable other researchers to reliably repeat the experimental work described.

D. ROC CURVE-THRESHOLDING TECHNIQUE
The Receiver Operating Characteristic (ROC) curve illustrates the performance capability of a classifier at various threshold settings. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [53] (Fig.1). The TPR and FPR are also known as sensitivity and probability of false alarm, respectively. The FPR can be calculated as (1 − specificity). TPR is on the y-axis and FPR is on the x-axis. The ROC curve is widely used to evaluate the performance of FR models, as it is a known classification task. In this case, the two classes are positive pairs - pair samples from the same identity (PP) and negative pairs - pair samples from two distinct identities (NP).

![FIGURE 1. A typical ROC curve, TP vs. FP rate at different classification thresholds [53], [54].](image1.png)

Also, it is common to derive a threshold from a relevant ROC curve and use this as a threshold for face recognition. Given two embeddings (numerical vectors), as the output of a face recognition model, representing two face samples, a score can be obtained representing the identity similarity of the two samples. This score is compared against the FR threshold and this comparison determines if the two samples have or not the same identity. The workflow of the thresholding technique described is illustrated in Fig.2. The threshold corresponds to an FPR value, which depending on the FPR value, makes the FR threshold more or less strict. In this work, the ROC curves and the thresholding technique, are the base of the Methodology created to answer the Research Questions presented in the Introduction.

![FIGURE 2. The workflow of Thresholding Technique.](image2.png)

IV. METHODOLOGY
In this work, we take two different approaches to measure the identity uniqueness of a set of synthetic data samples. In the
first approach, ROC curves are computed and compared to determine the identity uniqueness. This approach can be applied to determine the identity uniqueness of a synthetic dataset compared to the original set of seed data samples (Sy vs Se). It can also be applied to determine the uniqueness among the synthetic samples (Sy vs Sy).

In the second approach, a thresholding technique is used to calculate a new metric which allows making similar determinations of uniqueness between synthetic and seed data, and also among the synthetic samples. In this approach, a pair of samples are compared to an FR threshold to determine their identity similarity. As an extension, the second approach enables us to quantify the number of unique samples in a generated synthetic set of samples when compared with either the seed dataset or within itself. This can be used to measure the ability of a GAN model to generate synthetic data with unique identities.

This section is structured as follows: the section IV-A, explains the FR model alongside with how the synthetic samples are obtained. In section IV-B, the approach using ROC curves is described and finally, in IV-C the Thresholding Technique is explained in detail.

A. GENERATED SYNTHETIC DATA AND FACE RECOGNITION MODEL

For experimental purposes each generated synthetic dataset should meet several criteria: (i) the same parameters are used when generating synthetic facial samples (e.g., truncation psi); (ii) all generated samples are tested to ensure they are detected as a face; this ensures they can be correctly processed by the FR model which is an essential tool of this Methodology.

Also, as the Methodology depends on the FR model’s performance, it should be a state-of-art. An FR model is usually trained to output an embedding - numerical vector, which represents the input face sample. The embeddings are compared using a metric (e.g., Euclidean distance, cosine similarity, etc.), to get a score that represents the identity similarity. The FR model is optimized to output embeddings that for images of the same identity, the score computed shows high similarity compared to the score of images from different identities. The metric used to get the score from the two embeddings is selected based on the FR model’s implementation. The FR model in this Methodology is utilized by feeding it with face samples (real or synthetic) to get their corresponding embeddings. The embeddings are used to calculate the score for a pair representing their identity similarity, which is used to either compute the ROCs or for comparison against an FR threshold.

B. ROC CURVES COMPARISON

1) REFERENCE POINT ROC (REF-ROC)

To examine each case (Sy vs Se, Sy vs Sy), a ROC curve is computed, only using real samples. This ROC is used as a reference point (Ref-ROC) in the Methodology. This curve represents the statistical behavior of a dataset of real face samples. The Ref-ROC is compared with the ROC curves of synthetic samples and helps in understanding if the statistical distributions of the generated synthetic data samples match those of a real-world dataset. Thus, ROC curves are computed illustrating the statistical behavior of generated synthetic data for the various cases of interest – synthetic with the seed data (Sy vs Se) and synthetic with one another (Sy vs Sy) and compared against the Ref-ROC.

To compute a ROC curve the following procedure is followed. Initially, an equal number of positive and negative image pairs are created. A positive pair (PP) is when two face images have the same identity and in negative pairs (NP) have different identities. Using the corresponding embeddings (obtained by an FR model) of the image pairs, the scores are calculated and used to plot the ROC. The workflow of creating a Ref-ROC is given in Fig.3.

![FIGURE 3. The workflow of computing a Ref-ROC.](image-url)

2) IDENTITY UNIQUENESS BETWEEN SYNTHETIC AND SEED DATA (SY VS SE) - ROC CURVES COMPARISON

To examine the identity uniqueness between synthetic and the seed data, a ROC curve is computed using both synthetic and seed data (Sy-Se-ROC) and compared against the Ref-ROC curve. For the Sy-Se-ROC curve, the PPs remain the same as the PPs used in the Ref-ROC, but the NPs are different. The NP consist of pairing the generated synthetic data with the seed data (real face samples). The scores of the pairs (PPs, NPs) are computed using the corresponding embeddings and finally used to compute the Sy-Se-ROC. The workflow of computing a Sy-Se-ROC is given in Fig.4.

When the Sy-Se-ROC is compared against the Ref-ROC, the only difference between these two sets of ROCs (Sy-Se-ROC, Ref-ROC), as both use the same FR model and PPs, are the NPs. The NPs from the Ref-ROC consists of real NPs (as their identity is known), while the NPs from the Sy-Se-ROC are generated synthetic data (without an identity label) paired with seed data and therefore treated as NPs. As a result, when comparing the two ROCs, the behavior/performance of the
NP consisting of synthetic data and seed data (from the Sy-Se-ROC) is compared against the NPs from the Ref-ROC (which are real/known NPs).

Now either the Sy-Se-ROC is below the Ref-ROCs in different parts of the plot, or the Sy-Se-ROC is at the same or higher levels than the Ref-ROC. At higher or equal levels, we can conclude that that the probability of having a false positive from the NPs of the Sy-Se-ROC is the same or lower than that from the NPs of the Ref-ROC. In this case, the identity uniqueness between the synthetic data and the seed data is equal or higher as the one in real samples from different identities. Thus, showing that the generated synthetic data are unique when compared with the seed data in terms of identity, which is desirable.

When the Sy-Se-ROC is below the Ref-ROC, the probability of having a false positive from the NPs of the Sy-Se-ROC is higher than that of the NPs of the Ref-ROC. In this case, the identity uniqueness between the generated synthetic data and the seed data is lower compared to that of real samples, concluding that the generated data samples are not unique in terms of identity.

C. THRESHOLING TECHNIQUE AND UNIQUENESS METRICS

A common approach used in face recognition/verification to determine if two face samples are classified as having a similar identity or not is by using a thresholding technique. Using the embeddings of the two face samples, a score is obtained and compared against an FR threshold (Fig.2). This technique is followed to determine the identity uniqueness for a set of synthetic data for each case (Sy vs Se, Sy vs Sy).

1) FACE RECOGNITION THRESHOLD-SELECTION

The FR threshold used to implement this approach is representative of the statistical behavior of a dataset of real face samples, following the same reasoning as in section IV-B-1. Therefore, is derived from the Ref-ROC curve. Also, as described in section III-D, a threshold is derived from a ROC curve corresponds to an FPR value. The statistical meaning is that for a threshold corresponding to an FPR value, e.g., FPR=1e-05 means that statistically, 1 false positive is expected in every 100k comparisons.

2) IDENTITY UNIQUENESS BETWEEN SYNTHETIC AND SEED DATA (SY VS SE) - THRESHOLDING TECHNIQUE

In this case, the identity uniqueness between the generated synthetic data and seed data is examined (Sy vs Se), using the Thresholding Technique. All the generated synthetic data are paired with all the seed data. These pairs are considered NPs,
as the generated synthetic data don’t have an identity label. For all the pairs, their score is calculated through the corresponding embeddings. In continuance, the score is compared against the FR threshold to determine if the samples of the pair are classified as having a similar identity or not. The described workflow is illustrated in Fig 6. After calculating the metrics introduced in sections IV-C-4 and IV-C-5, the identity uniqueness of the generated synthetic data when compared with their seed data is determined.

3) IDENTITY UNIQUENESS AMONG THE SYNTHETIC DATA (SY VS SY) - THRESHOLDING TECHNIQUE

In this case, the identity uniqueness among the generated synthetic data is examined (Sy vs Sy), using the Thresholding Technique. The procedure is similar as in section IV-C-2, but as the identity uniqueness among the generated synthetic data is examined, the way that the pairs are created differ. In this case, all the generated synthetic data are paired with each other. These pairs are also treated as NPs. For all the pairs, their score is calculated through the corresponding embeddings. Then the score is compared against the FR threshold to determine if the generated synthetic samples of the pair are classified as having a similar identity or not. The described workflow is illustrated in Fig.7. The identity uniqueness is determined using the metrics introduced in the sections IV-C-4 and IV-C-5.

4) RATIO OF EXPECTED FALSE POSITIVES (REFP)

The number of pairs created for this thresholding technique is the number of comparisons that are made (NoC). The number of pairs that are classified as being similar in terms of identity, based on their score comparison with the FR threshold, is quantified (NoP). Due to the statistical meaning of FPR given in section IV-C-1, when a large number of comparisons is conducted, it is expected to have a number of pairs which, are classified as having a similar identity, but these might be statistically false positives.

As the generated synthetic data don’t have an identity label, to examine if there is actually an identity similarity (being classified as having a similar identity) between the samples of these pairs in each case (Sy vs Se, Sy vs Sy), the NoP, for the selected threshold corresponding to an FPR value, is compared with the number of expected statistical false positives on this FPR value.

If the NoP is at the same levels or lower than the expected statistical false positives on the FPR value, then it can be concluded that the synthetic samples are at the same or higher levels of identity uniqueness as the real samples from different identities, showing that the generated samples are unique in terms of identity for the case examined (Sy vs Se, Sy vs Sy). But if the NoP is higher, then the identity uniqueness of the generated synthetic data samples (for the examined case) is lower, concluding that the generated samples are not unique in terms of identity.

To account for different sizes of synthetic datasets, and thus different NoCs and different thresholds corresponding to different FPR values, the Ratio of Expected False Positive (REFP) is introduced and defined as:

\[
REFP = \left( \frac{NoP}{NoC} \right) / FPR
\]

The number of pairs in which its samples are classified as having a similar identity (NoP) is divided to the number of comparisons (NoC). This, in turn, is normalized by dividing it by the FPR value of the selected threshold. The REFP shows how many times higher or lower is the NoP, compared to the expected false positives, for the given FPR value, based on the comparisons conducted and therefore a lower value is desired showing a better performance.

If the REFP, is lower/equal or very close to 1, then the synthetic dataset for the case examined (Sy vs Se, Sy vs Sy) has a similar or higher identity uniqueness than real datasets. If the REFP is higher than 1, then the synthetic data samples are characterized by a lower level of uniqueness for the examined case. The REFP is a useful metric to understand if the GAN model is able to generate unique synthetic data for each examined case.
5) NUMBER OF UNIQUE SYNTHETIC DATA SAMPLES

A metric that can be used to evaluate the ability of a GAN model to generate unique synthetic data is to quantify these synthetic samples with a unique identity. This is possible as in the Thresholding Technique, all possible pair combinations have been taken into consideration.

Therefore, in the case where the synthetic data are not unique in terms of identity when compared to the seed data (REFP higher than 1), then calculating the number of synthetic samples with a unique identity is straightforward. We just subtract the number of synthetic samples that are classified at least once as having a similar identity with a seed sample from the total size of the synthetic dataset. This is defined as the number of unique samples (NoU) in a generated synthetic dataset when compared to their seed data.

In the case that the synthetic data show a lack of uniqueness (REFP higher than 1) when compared to one another, it is not straightforward to determine the samples with a unique identity as the identity similarities are entangled with other samples. But using a graph theory approach, it is feasible to determine the maximum number of unique identities in a generated synthetic dataset. The idea is to start by determining the most connected sample and remove this from the dataset. By iteratively removing the sample with the highest number of similarities/connections until no samples with a similarity/connection remain, it is possible to determine the number of unique data samples in a generated batch of synthetic data samples. This is defined as the number of unique samples (NoU) within a generated synthetic dataset. The procedure is given in Fig.8. In Appendix 1, the procedure is described through Algorithm 1 and an example where it is applied is given in Fig.21.

![FIGURE 8. The workflow of calculating the NoU within a synthetic set of samples (Sy vs Sy).](image)

Note that to compare the performance of different models, in the task of generating unique, in terms of identity, synthetic samples for each case, the NoU is used, as it shows the number of unique samples in a synthetic set while the REFP shows the number of pairs that have an identity similarity. Although as mentioned for the NoU to be computed, the REFP has to be higher than 1, showing lower level of uniqueness.

V. IMPLEMENTATION AND EXPERIMENTS

In this section, a series of experiments are presented based on the two approaches of the proposed Methodology.

A. SYNTHETIC DATASETS

In this work, the synthetic samples are generated from StyleGAN [3] and used as a basis for the experiments presented in this section. Three different StyleGAN models are used to build these synthetic datasets. Two official StyleGAN models from NVIDIA are used [3]. These are trained on FFHQ [3] and CelebA-HQ [48], respectively, at a 1024 × 1024 resolution. Also, a StyleGAN model trained on the CelebA [47] dataset at 256 × 256 resolution is used [32]. This enables the outcomes of this study to be validated across several different variants of the StyleGAN model so that the results are not biased to a specific seed training dataset. Each seed dataset has a different number of data samples and identities.

In addition, when generating data, it is possible to generate samples with different truncation psi values, as explained in section III-A. These experiments use three different values for this variable - 0.5, 0.7, and 1.0, so that the effects of varying this key parameter can be understood. Using these values allows the use of publicly available synthetic datasets provided from the StyleGAN model trained on FFHQ, allowing these experiments to be easily replicated using these samples [3]. The generated synthetic data are divided into several datasets of 20k samples used in this work.

Instructions on how to generate the same synthetic data samples for each set that is used in this work can be found in. In total, 9 sets of 20k generated synthetic datasets are used in this work. In order to refer to a dataset of synthetic samples, the following naming convention is used: Sy – name of seed dataset - truncation psi value (e.g., Sy-FFHQ-0.5). The generated synthetic datasets are given in Table 1 along with information of their corresponding seed dataset (e.g., number of samples and identities). Finally, as mentioned in IV-A, all the generated face samples used are “face detectable”.

| Synthetic (Sy) DB | Seed (Se) DB | Truncation psi value | # Se | # IDs in Se |
|------------------|--------------|----------------------|-------|-------------|
| Sy-FFHQ-0.5      | FFHQ         | 0.5                  | 70k   | 70k         |
| Sy-FFHQ-0.7      | FFHQ         | 0.7                  | 70k   | 70k         |
| Sy-FFHQ-1        | FFHQ         | 1                    | 70k   | 70k         |
| Sy-CelebA-HQ-0.5 | CelebA-HQ    | 0.5                  | 30k   | 6k          |
| Sy-CelebA-HQ-0.7 | CelebA-HQ    | 0.7                  | 30k   | 6k          |
| Sy-CelebA-HQ-1   | CelebA-HQ    | 1                    | 30k   | 6k          |
| Sy-CelebA-0.5    | CelebA       | 0.5                  | 200k  | 10k         |
| Sy-CelebA-0.7    | CelebA       | 0.7                  | 200k  | 10k         |
| Sy-CelebA-1      | CelebA       | 1                    | 200k  | 10k         |

B. FACE RECOGNITION MODEL

The FR tool used in this work is the ArcFace FR model [28]. The selected FR model takes a face image as an input and outputs a 512-embedding. The face samples are pre-processed

3https://github.com/NVlabs/stylegan
before passing to the FR model. The pre-processing includes a face detector, cropping the detected area and resizing to the required input size of the FR network. In this work, the same procedure as advised by the authors of ArcFace is followed before the face samples are fed to the network. The MTNN [33] is used to detect and crop the face samples and to validate if generated data samples are recognizable as a face. The detected area is cropped and resized to 112 × 112, using bilinear interpolation, before passing to the ArcFace recognition network which calculates the 512-embedding corresponding to a facial sample. Finally, the cosine similarity is used to compute the score representing the identity similarity when comparing two embeddings. The model and weights of ArcFace used in this work can be found in.

C. ROC CURVES COMPARISON

In this section, the ROC curves Comparison approach is implemented using the generated synthetic datasets (Table 1). Initially, the Ref-ROCs are computed and using the generated synthetic datasets, the Sy-Se-ROCs and the Sy-ROCs are computed. The identity uniqueness between samples from the generated synthetic datasets with the samples from their corresponding seed datasets is examined by comparing the Sy-Se-ROC with the Ref-ROC. Also, the identity uniqueness among the samples from each generated synthetic dataset is examined by comparing the Sy-ROC with the Ref-ROC. In both cases, the influence of truncation psi and the performance of the different models are also explored.

1) COMPUTING THE REF-ROC

For this work, three Ref-ROCs are computed with NPs taken from different datasets. In this way, the results are not specific to a single dataset or Ref-ROC curve. The PPs are all the possible PPs that can be formulated from the CelebA dataset, in total 2.5M. The NPs for these three Ref-ROCs are formulated by combining a sample from the CelebA dataset [47], with samples from CelebA, LFW [45] and CasiaWebFaces [46] datasets respectively as summarized in Table 2. In each case, 2.5M NPs are created to match the number of PPs.

| Name of Ref-ROC | PP Dataset | NP Datasets |
|-----------------|------------|-------------|
| CelebA          | CelebA     | CelebA      |
| LFW             | CelebA     | LFW         |
| CasiaWebFaces   | CelebA     | CasiaWebFaces |

2) DETERMINING UNIQUENESS - SY VS SE - ROC CURVES COMPARISONS

In order to determine the identity uniqueness of each synthetic dataset with their seed dataset, a corresponding Sy-Se-ROC is computed which is described in section IV-B-2 and shown in Fig.4. It uses the same 2.5M PPs used in the Ref-ROC and 2.5M NPs of synthetic data samples paired with seed data samples. In Table 3, these different combinations are listed.

| Name of Sy-Se-ROC | PP Dataset | NP Datasets |
|-------------------|------------|-------------|
| Sy-Se-ROC         | CelebA     | CelebA      |
| FFHQ-0.5          | CelebA     | FFHQ        |
| FFHQ-0.7          | CelebA     | FFHQ        |
| FFHQ-1           | CelebA     | FFHQ        |

In Fig.9 and 10, the Sy-Se-ROCs, start and remain at similar levels or above the Ref-ROCs. This indicates that the

![FIGURE 9.](image-url) The Sy-Se-ROCs from StyleGAN-CelebA (Table 3) compared against the Ref-ROCs to determine the identity uniqueness between the synthetic and the seed samples.

![FIGURE 10.](image-url) The Sy-Se-ROCs from StyleGAN-CelebA-HQ (Table 3) compared against the Ref-ROCs to determine the identity uniqueness between the synthetic and the seed samples.
identity uniqueness between the synthetic data and the seed data is equal or higher as the one in real samples with different identities. Thus, showing that the generated synthetic data from these models (StyleGAN-CelebA and StyleGAN-CelebA-HQ) are unique when compared with their seed data in terms of identity.

In Fig.11, the Sy-Se-ROCs using the synthetic datasets generated from the StyleGAN-FFHQ model are compared against the Ref-ROCs. All the Sy-Se-ROCs start and remain below the Ref-ROCs, indicating that the generated synthetic data from the StyleGAN-FFHQ are not unique when compared with their seed data samples. However, when a selection of the NPs with high similarity scores was examined it was clear that the high-scoring pairs consisted of face samples of infants or young children. The FFHQ dataset is unique among our selected datasets as it is the only dataset to contain such samples. As the reference model of ArcFace was not trained with samples of kids/babies and therefore it isn’t robust in distinguishing such samples, which explains the results of Fig 11. Thus in Fig.12, the same Sy-Se-ROCs as in Fig.11 are presented but with the manual removal of several data pairs that consist of infants or young children. It is clear that these remain at similar levels or above the Ref-ROC and the behavior is now similar to Fig.9 and 10. This indicates that the identity uniqueness between the synthetic data and the seed data is equal or higher as the one in real samples, showing that the generated synthetic samples from StyleGAN-FFHQ are unique when compared with their seed data in terms of identity.

In order to examine the influence of the truncation psi value in the identity uniqueness between the generated synthetic data and the seed data, the Sy-Se-ROCs of each figure (Fig. 9,10,12) are compared with each other. From this comparison, it is observed that all the Sy-Se-ROCs perform similarly with only marginal differences. This suggests that the value of the truncation psi parameter does not influence the identity uniqueness of generated synthetic data samples with respect to their seed dataset.

Finally, in Fig.13, all Sy-Se-ROCs are presented together to determine which StyleGAN model is better at generating synthetic data which are unique in terms of identity when compared with their seed data. From this comparison, the StyleGAN-FFHQ shows a better performance, followed by the StyleGAN-CelebA and StyleGAN-CelebA-HQ, respectively but with only marginal differences between them.

In summary, it can be concluded that StyleGAN is very effective at generating synthetic data samples that are well distinguished (unique in terms of identity) from the original set of data samples used to train the GAN. The next challenge is to understand how unique the synthetic data samples are when compared with other synthetic samples.

3) DETERMINING UNIQUENESS - SY VS SY - ROC CURVES COMPARISONS

In order to determine the identity uniqueness within each synthetic dataset, a corresponding Sy-ROC is computed which
is described in section IV-B-3 and shown in Fig 5. It uses the same 2.5M PPs used in the Ref-ROCs and 2.5M NPs of synthetic samples paired with one another. In table 4, these different combinations are listed.

**TABLE 4.** The synthetic datasets used to compute each sy-roc curve, as shown in Fig.5.

| Name of Sy-ROC | PP Dataset | NP Dataset |
|----------------|------------|------------|
| CelebA-0.5     | CelebA     | Sy-CelebA-0.5 |
| CelebA-0.7     | CelebA     | Sy-CelebA-0.7 |
| CelebA-1       | CelebA     | Sy-CelebA-1  |
| CelebA-HQ-0.5  | CelebA     | Sy-CelebA-HQ-0.5 |
| CelebA-HQ-0.7  | CelebA     | Sy-CelebA-HQ-0.7 |
| FFHQ-0.5       | CelebA     | Sy-FFHQ-0.5  |
| FFHQ-0.7       | CelebA     | Sy-FFHQ-0.7  |
| FFHQ-1         | CelebA     | Sy-FFHQ-1    |

In Fig. 14-16, the Sy-ROCs are compared against the Ref-ROCs and start and remain below them. As explained in the corresponding methodology of this experiment in section IV-B-3, this shows that there is a lower statistical identity uniqueness among the samples of a synthetic dataset compared to that of real samples. Concluding that, when generating synthetic samples with any model used in this work (StyleGAN-CelebA, StyleGAN-CelebA-HQ, StyleGAN-FFHQ), not all the generated synthetic samples have a unique identity.

In order to examine the influence of the truncation psi value in the identity uniqueness among the generated synthetic data of a synthetic dataset, the Sy-ROCs of each figure (Fig. 14-16) are compared between them. The Sy-ROCs corresponding to a truncation psi value closer to 0, has a lower performance compared to the others, in all figures. The lower the performance of a Sy-ROC, the lower the number of synthetic samples that have a unique identity in a synthetic dataset. Concluding that the truncation psi influences the identity uniqueness of the generated synthetic samples, showing that when generating synthetic samples, the closer the truncation psi value is to 0, the lower is the identity uniqueness in the set of synthetic samples. This is an anticipated result that is validated through our experiments. When the truncation psi is closer to 0, the latent space chosen to generate the image is more truncated (section III-A) and, therefore with a less overall variation which extends in less variation (less uniqueness) in the identity feature of the generated synthetic samples.

Finally, in Fig. 17, all the Sy-ROCs are compared between them to examine which model performs better in generating synthetic samples with unique identities. We compare the models between them by comparing their corresponding Sy-ROC with the same truncation psi value between them. This is performed, for the comparisons to be fair and consistent, as the truncation psi influences the identity uniqueness among the generated synthetic samples. (e.g., the Sy-ROC-CelebA-0.5, the Sy-ROC-CelebA-HQ-0.5, and the Sy -ROC- FFHQ-0.5 are compared between them).
Comparing the Sy-ROCs as explained, from Fig. 17, it is illustrated that the Sy-ROCs corresponding to the StyleGAN-FFHQ is performing overall better followed in performance by the StyleGAN-CelebA and StyleGAN-CelebA-HQ, respectively, for all the different truncation psi values. The higher the performance of a Sy-ROC, the higher the number of generated synthetic samples that have a unique identity in the synthetic dataset. As a result, it is shown that when using the same truncation psi value, the StyleGAN-FFHQ model performs the best in generating synthetic samples with a unique identity followed in performance by the StyleGAN-CelebA and the StyleGAN-CelebA-HQ respectively.

D. THRESHOLDING TECHNIQUE

In this section, the Thresholding Technique is implemented using the generated synthetic datasets (Table 1). Initially, the FR threshold selection is described. Then, the identity uniqueness between the samples from the synthetic datasets with the samples from their corresponding seed datasets and as well the identity uniqueness among the samples from each synthetic dataset is examined. In both cases, the influence of truncation psi and the performance of the different models is explored using the uniqueness metrics presented in the Methodology (section IV-C).

1) FACE RECOGNITION THRESHOLD-SELECTION

To implement the Thresholding Technique as described in section IV-C-1, the FR threshold is derived from the Ref-ROC curve. The thresholds are derived from the Ref-ROC-CelebA (Table 2), as two out of three StyleGAN models used in this work are trained on the same distribution (StyleGAN-CelebA, StyleGAN-CelebA-HQ). However, the threshold could have been derived from any other Ref-ROC. For this work, three thresholds are selected corresponding to different FRP values. The threshold values are 0.483, 0.428, 0.3815 corresponding to the FPR values of: 4.46e-07 (the closest to 1e-07), 1.33e-06 (the closest to 1e-06) and 1e-05. Three thresholds are selected to examine how the identity uniqueness in each case (Sy vs Se, Sy vs Sy) is influenced as the threshold is less or more strict. As mentioned in section V-B, the score that represents the identity similarity between two samples through their corresponding embeddings is the cosine similarity. As for the cosine similarity metric, a higher score shows a higher similarity in terms of identity, in the comparison of the pair’s score with the FR threshold, if the score is above the FR threshold then the samples of the pair are classified as having a similar identity.

2) DETERMINING UNIQUENESS - SY VS SE - THRESHOLDING TECHNIQUE

In this experiment, the uniqueness of the generated synthetic data with the samples from the seed data in terms of identity is examined following the Threshold Technique described in section IV-C-2. The generated synthetic datasets are compared and with their corresponding seed data (Table 5). For each synthetic dataset the number of pairs that their score is above the determined FR threshold (NoP) is calculated. As the number of comparisons (NoC) is also known along with the FPR value, the REFP described in (1) is calculated for each synthetic dataset and is given in Table 6. For more numerical details on the NoP and NoC, see Table 15, in Appendix B.

| Name of Comparison | Synthetic DB | Seed DB |
|-------------------|--------------|--------|
| Sy-Se-CelebA-0.5  | Sy-CelebA-0.5| CelebA |
| Sy-Se-CelebA-0.7  | Sy-CelebA-0.7| CelebA |
| Sy-Se-CelebA-1    | Sy-CelebA-1  | CelebA |
| Sy-Se-CelebA-HQ-0.5| Sy-CelebA-HQ-0.5| CelebA-HQ|
| Sy-Se-CelebA-HQ-0.7| Sy-CelebA-HQ-0.7| CelebA-HQ|
| Sy-Se-FFHQ-0.5    | Sy-FFHQ-0.5  | FFHQ   |
| Sy-Se-FFHQ-0.7    | Sy-FFHQ-0.7  | FFHQ   |
| Sy-Se-FFHQ-1     | Sy-FFHQ-1    | FFHQ   |

From Table 6, the REFP for the synthetic datasets generated from the models StyleGAN-CelebA and StyleGAN-CelebA-HQ, in all the different FR thresholds, is lower or just marginally higher than 1. This shows that the synthetic data generated from the StyleGAN-CelebA and
StyleGAN-CelebA-HQ, are unique with respect to their corresponding seed samples.

For StyleGAN-FFHQ, the REFP is significantly higher than 1, but as discussed in V-C-2, this behavior is due to the presence of data samples of infants and young children. As to the best of our knowledge there isn’t a direct way to eliminate all the pairs consisting of only infants/young children. Also, if this is conducted the ability of the model-StyleGAN-FFHQ in generating synthetic samples would be decreased and the comparisons won’t be fair; therefore the REFP can’t be re-calculated. Eliminating pair samples consisting of only infants/young children solves this problem, as discussed and demonstrated in section V-C-2, where it is concluded that the generated synthetic data from the StyleGAN-FFHQ model are unique when compared with the samples from their seed dataset (FFHQ).

Regarding the influence of the truncation psi in the identity uniqueness between the synthetic samples and the samples from their seed datasets, the REFP values of each set are compared in Table 6. For the StyleGAN-CelebA and StyleGAN-CelebA-HQ, at a particular threshold, the values are reasonably consistent and lower or just marginally higher than 1, and it is concluded that truncation psi value does not influence on the identity uniqueness, a similar conclusion to that reached in section V-C-2. The behavior of StyleGAN-FFHQ doesn’t show this consistency, but this may be explained by the unpredictable effects of data samples of young children, where when omitted for the StyleGAN-FFHQ, it is shown that the truncation psi does not influence the uniqueness.

As the REFP is lower or just marginally higher than 1 for all the synthetic datasets showing that the generated synthetic samples have a unique identity when compared with samples from their seed dataset, the NoU can’t be calculated. Therefore, to compare the performance of the different synthetic datasets, the REFP is used. The synthetic datasets that are generated with the StyleGAN-CelebA and StyleGAN-CelebA-HQ, at a particular threshold, the values are reasonably consistent and lower or just marginally higher than 1, and it is concluded that truncation psi value does not influence on the identity uniqueness, a similar conclusion to that reached in section V-C-2. The behavior of StyleGAN-FFHQ doesn’t show this consistency, but this may be explained by the unpredictable effects of data samples of young children, where when omitted for the StyleGAN-FFHQ, it is shown that the truncation psi does not influence the uniqueness.

3) DETERMINING UNIQUENESS - SY VS SY - THRESHOLDING TECHNIQUE

In this experiment, the identity uniqueness among the generated synthetic data is examined following the Threshold Technique described in section IV-C-3. For each generated synthetic dataset, their samples are cross-compared (Table 7) to calculate the number of pairs with similar identities (NoP). Next, using the number of comparisons (NoC), the selected FPR, the REFP of (1) is calculated for each synthetic dataset and listed in Table 8. For more numerical details on the NoP and NoC, see table XVI, Appendix C.

For all the synthetic sets, and across all threshold settings, the REFP is significantly above 1 (Table 8). It is clear that when generating synthetic data samples from the models used in this work, not all synthetic samples have a unique identity. In fact, even our best result shows that the REFP rate is about 36 times higher than would be expected in a real-world dataset.

Regarding the influence of the truncation psi in the identity uniqueness of the generated samples, it can be seen that the REFP is reduced at higher values of this variable. It can be concluded that a larger truncation psi value increases the identity uniqueness of the generated synthetic samples, with the best results obtained when this variable is at value of 1.0.

The REFP is above the 1 for all the synthetic sets showing that not all the generated synthetic samples have a unique identity. In such a case, as discussed in section IV-C-5, the number of unique samples (NoU) can be calculated. It is also useful to provide the percentage of synthetic data with a unique identity in the synthetic dataset. These values are given in Table 9.

The NoU is used to compare the performance of different models in the task of generating face samples with a unique identity. The NoUs for each model with the same truncation psi value are compared between them. This is performed as the truncation psi influences the identity uniqueness of the generated synthetic samples. Comparing the NoU, in the different threshold settings, it is shown that the highest NoU is from the synthetic datasets generated with the StyleGAN-FFHQ model, showing the best performance, followed by the StyleGAN-CelebA. Finally, last in
performance is the StyleGAN-CelebA-HQ, in the task of generating synthetic samples with unique identities.

The NoU also shows that even in the strictest FR threshold of this work (FPR=4.46e-07, Threshold=0.4823), and using the full variation of the latent space (truncation psi=1), in the synthetic datasets, 80-90% of the samples have a unique identity. More interestingly, when the FR threshold becomes less strict (FPR=1e-05, Threshold=0.3815) and using a truncation psi value closer to 0 (truncation psi =0.5), while ensures better quality for the output image, the samples with a unique identity in the datasets decrease dramatically to 7-9%. This shows the necessity of investigating the variation in the identity attribute of the generated synthetic samples, as in some cases only a small number of samples are unique (in terms of identity). The samples with a unique identity (Table 9) from each synthetic dataset and for the different thresholds, which can form a seed synthetic face dataset with distinct identities are made available through the IEEEDataPort accompanying this article. Also, these can be found in.

E. VISUAL EXAMPLES

In this section, some qualitative examples are provided, to visually demonstrate how the proposed Methodology removes the most connected data samples to provide a unique set of synthetic identities.

The thresholding technique is able to locate synthetic pair samples that have similar identities. In Fig.18, several such synthetic pair samples with a high similarity score are shown. Visual inspection shows how similar these data samples are and serves to illustrate the key challenge to achieve unique identity seed samples in order to creating a valid synthetic dataset.

In Fig.19, state A, a cluster of synthetic samples is illustrated. These samples are interlinked by high similarity scores indicating they are not unique from one another. In Table 10, the identity similarity scores between these data samples is shown. The FR threshold in this example is: 0.48 and it can be seen from the highlighted scores that none of these identities can be considered unique. The number of connections (con) that each sample has is also given in Table 10. By applying our graph theory approach (section IV-C-5), the most connected face sample, Sy-1, can be removed. The four remaining samples, shown as State B, do not have an identity similarity and thus can be considered as unique seed identities.

![FIGURE 18. The synthetic pair samples have a similarity score above the FR threshold of this work (0.48). Thus, the synthetic samples have an identity similarity and therefore are not unique. The similarity score for the pairs is: A=0.72, B=0.70, C=0.75, D=0.74.](image1)

![FIGURE 19. A cluster of synthetic samples which have an identity similarity between them (state A). The graph theory technique of section IV-C-5 removes the sample with the most identity connections resulting in four unique identities (state B). The corresponding similarity scores, and number of connections are given in Tables 10 and 11.](image2)

| Samples | Sy-1 | Sy-2 | Sy-3 | Sy-4 | Sy-5 | Con |
|---------|------|------|------|------|------|-----|
| Sy-1    | -    | 0.62 | 0.53 | 0.52 | 0.49 | 4   |
| Sy-2    | 0.62 | -    | 0.39 | 0.30 | 0.36 | 1   |
| Sy-3    | 0.53 | 0.39 | -    | 0.28 | 0.23 | 1   |
| Sy-4    | 0.52 | 0.30 | 0.28 | -    | 0.29 | 1   |
| Sy-5    | 0.49 | 0.36 | 0.23 | 0.29 | -    | 1   |

\[5\text{https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Synthetic_Face_Datasets}\]
TABLE 11. The identity similarity score between the samples (Fig.19-state B) and the number of connections of each sample.

| Samples | Sy-2 | Sy-3 | Sy-4 | Sy-5 | Con |
|---------|------|------|------|------|-----|
| Sy-2    | -    | 0.39 | 0.30 | 0.36 | 0   |
| Sy-3    | 0.39 | -    | 0.28 | 0.23 | 0   |
| Sy-4    | 0.30 | 0.28 | -    | 0.29 | 0   |
| Sy-5    | 0.36 | 0.23 | 0.29 | -    | 0   |

FIGURE 20. A cluster of synthetic samples which have an identity similarity between them (state A). The graph theory technique of section IV-C-5 removes the samples with the most identity connections in two steps, resulting in four unique identities (state C). The corresponding similarity scores, and number of connections are given in Tables 12-14.

Another similar but more complicated example is given in Fig.20, with the Tables 12-14 showing the identity similarity scores for each state (A-C) in Fig 20, respectively. From Table 12, it is shown that the Sy-1 (Fig.20, state A) is the sample with the most connections (5) and therefore eliminated. After the Sy-1 is removed, in table 13, the Sy-2 (Fig.20, state B), is the sample with the most connections (2) and therefore also removed. Table 14, which corresponds to state C from Fig.20, shows that the remaining samples do not have a connection between them and therefore these samples are unique, in terms of their identity. These are simplified examples of the graph theory technique (section IV-C-5) applied to a small subset of a synthetic dataset. As discussed, and illustrated in section V-D, when used to the entire synthetic dataset, it allows us to identify and quantify the samples with a unique identity within a synthetic dataset. This can be used to measure the ability of GAN in generating synthetic samples with a unique identity or select these unique samples to be used in further research.

F. COMPUTATIONAL ASPECTS OF ROC CURVES COMPARISON AND THRESHOLDING TECHNIQUE

The ROC curve Comparison approach (section IV-B), uses a lower number of comparisons compared to the Thresholding Technique (section IV-C). This is because a limited number of random pairs are selected as representative of the statistical properties of the underlying data distribution.

The Thresholding Technique can provide more precise results for a generated set of synthetic data and quantify the uniqueness of these samples with either the seed dataset or the other synthetic data samples in the generated dataset. This also allows the REFP metric to be calculated along with the NoU which as shown can be used to measure the number of synthetic data with a unique identity. However, the number of comparisons can quickly become a limiting factor - when using a dataset size of 20k synthetic samples with a corresponding seed dataset of 200k then 4B comparisons are needed to examine the identity uniqueness between the two sets and 200M comparisons are required to examine the uniqueness within the synthetic dataset, which can take up to a day to calculate. On the other hand, the ROC based approach requires only 5M comparisons in each case.

VI. CONCLUSION AND FUTURE WORK

In this work, a Methodology is presented with two different approaches that enable to answer the following Research Questions posed in this work:

1) Are the synthetic data samples unique when compared with the original seed data used to train the GAN model?

2) Are the synthetic data samples within a generated dataset unique when compared with one another?

3) Can we validate individual samples within a generated dataset to ensure that there is sufficient identity uniqueness to use as a synthetic data sample for further research?

In both approaches, the performance/behavior of real samples is used as a reference point and the performance/behavior of the generated synthetic data for each case (Sy vs Se, Sy vs Sy) is compared against it to answer the
Research Questions. In the first approach the performance/behavior of real samples and generated synthetic data for each examined case is illustrated through ROC curves, which are compared and examine the uniqueness of the generated synthetic data with the seed data and the uniqueness among the generated synthetic data, in terms of identity. In the second approach, a Thresholding Technique is implemented to determine the identity uniqueness for both cases. In this approach, the performance of real samples is illustrated through the FR thresholds that are selected and used to determine the similarity of the generated synthetic samples with their seed dataset and also among them, which reversely shows the identity uniqueness of the synthetic data in each case. Using this approach, the introduced metric REFP is calculated which answers the questions of this work. Also, through the Thresholding Technique approach, another metric can be calculated which is used to measure a model’s ability to generate samples with a unique identity and also identify these samples (NoU).

To answer the Research Questions, the two presented approaches are implemented using generated samples from three different StyleGAN [3] models using different settings (e.g. truncation psi value). In this way, the identity uniqueness is examined, for both cases, for several models and different truncation psi values. StyleGAN is selected as it represents the state-of-the-art GAN for the face generation task, although any GAN model trained on this task can be used to implement the Methodology proposed.

Given the extensive observations in section V, both approaches concluded to similar results. Both approaches concluded that the generated synthetic samples from any model used in this work are as unique in terms of identity with the samples from their corresponding seed data, as samples from different identities in a real dataset, which is desirable. Also generating samples with any truncation psi value does not influence the identity uniqueness between the generated synthetic and their seed data. Finally, all the models perform similarly in this task with small differences between them.

When comparing the synthetic samples with one another, both approaches concluded that using the models of this work, the generated samples are not as unique in terms of identity as samples from different identities in a real dataset. Also, it is shown that the truncation psi value influences the identity uniqueness of generated synthetic samples of a set, for any model used in this work. When generating samples with a truncation psi value closer to 0, the identity uniqueness in the synthetic datasets is lower in comparison when generating samples with a truncation psi value further from 0. Additionally, it is shown, that the model of StyleGAN-FFHQ, performs the best in generating synthetic samples with a unique identity when compared with each other, followed by the StyleGAN-CelebA and StyleGAN-CelebA-HQ. Finally, the NoU metric, which shows the ability of the models to generate unique synthetic samples, reveals that in some cases only 7-9% of the samples have a unique identity from a dataset of 20k generated synthetic samples.

In future work, this Methodology can be used to build synthetic facial datasets at scale by using a GAN to generate a seed dataset of facial data samples that are demonstrably unique in terms of their identity. Given such a seed dataset it would then be feasible to modify features (e.g. facial lighting, pose, and expression) of these seed samples to build large synthetic training datasets that could be used for FR purposes and other applications. Furthermore, using this methodology, it would be interesting to investigate and benchmark different GANs for the task of generating face samples with a unique identity. Also, future work will include a subjective evaluation which additionally will allow to further study the validity of the face recognition system in the examined cases.
Finally, utilizing core ideas from the proposed Methodology, a loss function can be created that can be used to train GAN models in order to maximize their ability to generate facial data samples with a unique identity.

APPENDIX A

Algorithm 1, below indicates the graph approach which allows to quantify the number of samples with a unique identity within a synthetic dataset, described in section IV-C-5 and in Fig 21, this is shown, as the Algorithm 1 is applied to a simplified graph.

![Figure 21. A graph containing several synthetic samples. The arrows show if two samples have an identity similarity between them and next to them is the number of connections of each sample. The Algorithm 1 is applied till the remaining samples do not have any connection between them which gives the NoU within a synthetic dataset.](image)

### TABLE 15.

The number of pairs with similar identity (NoP) using the synthetic dataset are given in Table 15.

| Name of Comparison | FPR/ Threshold | NoC  |
|--------------------|----------------|------|
| Sy-Se-CelebA-0.5   | 4.46e-07/0.4823| 27572| 4B   |
| Sy-Se-CelebA-0.7   | 4.46e-07/0.4823| 27572| 4B   |
| Sy-Se-CelebA-1     | 5.521          | 27771| 4B   |
| Sy-Se-CelebA-HQ-0.5| 3.657          | 7068 | 600M |
| Sy-Se-CelebA-HQ-0.7| 90             | 7068 | 600M |
| Sy-Se-FFHQ-0.5     | 3.176          | 61295| 1.4B |
| Sy-Se-FFHQ-0.7     | 14.688         | 161285| 1.4B |
| Sy-Se-FFHQ-1      | 40.71          | 17.88| 1.4B |

APPENDIX B

Following the Thresholding Technique described in section IV-C-2, the uniqueness of the synthetic datasets with the samples from the seed data in terms of identity is examined in section V-D-2 and the NoP and NoC for each synthetic dataset are given in Table 15.

### TABLE 16.

The number of pairs with similar identity (NoP) using the Thresholding Technique for the Sy vs Sy case and the number of comparisons (NoC).

| Name of Comparison | FPR/ Threshold | NoP  |
|--------------------|----------------|------|
| Sy-Sy-CelebA-0.5   | 4.46e-07/0.4823| 3.005.553| 6330.995| 200M |
| Sy-Sy-CelebA-0.7   | 3.176          | 339.832| 894.335| 200M |
| Sy-Sy-CelebA-1     | 5.521          | 22.921| 73.003| 200M |
| Sy-Sy-CelebA-HQ-0.5| 5.837          | 4.299.750| 8.119721| 200M |
| Sy-Sy-CelebA-HQ-0.7| 209.072       | 641.038| 1.499.136| 200M |
| Sy-Sy-FFHQ-0.5     | 13.004        | 50.335| 146.326| 200M |
| Sy-Sy-FFHQ-0.7     | 32.533        | 124.973| 365.651| 200M |
| Sy-Sy-FFHQ-1      | 11.445        | 34.431| 84.741| 200M |

APPENDIX C

Following the Thresholding Technique described in section IV-C-3, the uniqueness within the synthetic datasets, in terms of identity, is examined in section V-D-3, and the NoP and NoC for each synthetic dataset are given in Table 16.

### REFERENCES

1. D. Berthelot, T. Schumm, and L. Metz, “BEGAN: Boundary equilibrium generative adversarial networks,” 2017, arXiv:1703.10717. [Online]. Available: http://arxiv.org/abs/1703.10717
2. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, arXiv:1511.06434. [Online]. Available: http://arxiv.org/abs/1511.06434
3. T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4401–4410.
4. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
5. S. Bazaar, H. Javidinia, and P. Corcoran, “Latent space mapping for generation of object elements with corresponding data annotation,” Pattern Recognit. Lett., vol. 116, pp. 179–186, Dec. 2018.
6. H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs, “Deep single-image portrait relighting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7194–7202.
7. Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of GANs for semantic face editing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9243–9252.
8. A. Borji, “Pros and cons of GAN evaluation measures,” Comput. Vis. Image Underst., vol. 179, pp. 41–68, Feb. 2019.
9. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2234–2242.
10. W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, “Many paths to equilibrium: GANs do not need to decrease a divergence at every step,” 2017, arXiv:1710.08446. [Online]. Available: http://arxiv.org/abs/1710.08446
11. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2818–2828.
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
13. S. Gurumurthy, R. K. Sarvadevabhatla, and R. V. Babu, “DeLiGAN: Generative adversarial networks for diverse and limited data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 166–174.
14. X. Chen, “Improved techniques for training GANs,” 2016, arXiv:1612.012136. [Online]. Available: http://arxiv.org/abs/1612.012136
15. Z. Zhou, H. Cai, S. Hong, Y. Song, K. Ren, W. Zhang, Y. Yu, and J. Wang, “Activation maximization generative adversarial nets,” 2017, arXiv:1703.02000. [Online]. Available: http://arxiv.org/abs/1703.02000
16. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.
17. R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 649–666.
GABRIEL COSTACHE received the B.Sc. and M.Sc. degrees from the Faculty of Electronics and Telecommunications, Politehnica University of Bucharest, Romania, in 2003 and 2004, respectively, and the Ph.D. degree in image processing from the National University of Ireland Galway (NUIG), in 2006. He has been a part of FotoNation, since 2006. His current role is the Director of biometrics at FotoNation/Xperi, which develops technologies to process 2-D and 3-D imaging data.

PETER CORCORAN (Fellow, IEEE) holds the Personal Chair in electronic engineering at the College of Science and Engineering, National University of Ireland Galway. He is currently an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He was a Co-Founder in several start-up companies, notably FotoNation, now the Imaging Division of Xperi Corporation. He has over 600 technical publications and patents, over 100 peer-reviewed journal articles, 120 international conference papers, and a co-inventor of more than 300 granted U.S. patents. He is a member of the IEEE Consumer Electronics Society for over 25 years. He is the Editor-in-Chief and the Founding Editor of IEEE Consumer Electronics Magazine.

* * *