Title	Contributions to data augmentation techniques and synthetic data for training deep neural networks
Author(s)	Varkarakis, Viktor
Publication Date	2022-06-10
Publisher	NUI Galway
Item record	http://hdl.handle.net/10379/17194

Contributions to Data Augmentation Techniques and Synthetic Data for Training Deep Neural Networks

OÉ Gaillimh
NUI Galway

Viktor Varkarakis

College of Engineering and Informatics
National University of Ireland, Galway

This dissertation is submitted for the degree of
Doctor of Philosophy

Supervisor: Prof. Peter Corcoran   June 2022

"The only true wisdom is in knowing you know nothing. "

– Socrates

Table of contents

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

<div align="right">

Viktor Varkarakis

June 2022

</div>

Acknowledgements

In this section I would like to express my appreciation and to acknowledge almost every person who has helped me in any way during my PhD journey.

First of all, I would like to thank my supervisor and mentor, Peter Corcoran. Professor Corcoran was the one who initially suggested that I apply for a PhD position here in Galway and gave me eventually a Ph.D. position. I'll always be appreciative for the many opportunities you have given and continue to give me. You were always supportive and made me believe that I can accomplish the Ph.D journey. I can never get bored discussing with you and I wish we had the opportunity to go to more conferences together with you as I have amazing memories from the conference in LA that we attended together. Thank you again. You are amazing, not only as a supervisor but as a person and I learned a lot from you and I am still learning.

I would like to thank Professor Chris Dainty. It is rare to have the opportunity to interact with a scholar of your accomplishments and yet you always had time for anyone. I will always remember one of your teasing quotes: "There are 24hrs in a day", showing us that we can accomplish anything we want even in a very short period of time.

I would like to thank Gabriel Costache, who despite being a very busy manager in Fotonation/Xperi, always had the time to answer my questions, help me advance my research.

Alexandru Drimbarean, you were one of my first contacts in Fotonation and gave me the opportunity for an Erasmus+ placement in Fotonation. Without this opportunity I doubt I would be here today. Thank you.

Shejin Thavalengal, you were the one that helped me apply for the Ph.D. and mentor me. Thank you for your support and even when you left Ireland you still kept helping me.

Shabab Bazrafkan, you were an amazing mentor, taking me under your wing and helping me especially in the beginning of my PhD. It was a gift collaborating with you. Thank you for bringing your energy and smile everyday no matter the circumstances.

Hossein Javidnia, for your friendship and support during my PhD and for all the chicken nuggets we shared.

Abstract

In the recent years deep learning has become more and more popular and it is applied in a variety of fields, yielding outstanding results in different machine learning applications. Deep learning based solutions thrive when a large amount of data is available for a specific problem but data availability and preparation are the biggest bottlenecks in the deep learning pipelines. With the fast-changing technology environment, new unique problems arise daily. In order to realise solutions in many of these specific problem domains there is a growing need to build custom datasets that are tailored for a particular use case with matching ground truth data. Acquiring such datasets at the scale required for training with today's AI systems and subsequently annotating them with an accurate ground truth is challenging. Furthermore, with the recent introduction of GDPR and associated complications introduced, industry now faces additional challenges in the collection of training data that is linked to individual persons.

This dissertation focuses on ways to overcome the unavailability of real data and avoid the challenges that come with a data acquisition process. More specifically data augmentation techniques are proposed to overcome the unavailability of real data, improve performance and allow the use of low-complexity models, suitable for implementation in edge devices. Furthermore, the idea of using AI tools to build large synthetic datasets is considered as an alternative to real data samples. The first steps in order to build and incorporate synthetic datasets effectively into the deep learning training pipelines include: building AI tools, that will generate a large amount of new data and/or augment these data samples and also create methodologies and techniques to validate that the generate data behave like real ones and also measure whether their use is effective when incorporated in the training pipelines, with this dissertation contributing to both of these steps.

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AI   Artificial Intelligence

ANN   Artificial Neural Network

AR   Augmented Reality

CE   Consumer Electronics

CNN   Convolutional Neural Networks

CPU   Central Processing Unit

DL   Deep Learning

DNN   Deep Neural Networks

FR   Face Recognition

GAN   Generative Adversarial Network

GDPR   General Data Protection Regulations

GPU   Graphics Processing Unit

ML   Machine Learning

NIR   Near-Infrared

ROC   Receiver Operating Characteristic

SoA   State-of-Art

SVM   Support Vector Machines

VR   Virtual Reality

Chapter 1

Introduction

1.1 Introduction

The concept of artificial neural networks (ANNs) can be traced back to the 1960's [1]. Due to computational cost and a lack of understanding of how ANN work, simpler models that use task-specific handcrafted features such as Gabor filters and support vector machines (SVM) were a popular choice in the 1990s and 2000s [2]. In the last decade, due to a combination of the availability of massive data sets (Big Data), and the arrival of new graphics-processing-unit (GPU)-based hardware that enables these large data sets to be processed in reasonable timescales, triggered an exponential growth in research activity into the advanced training of ANNs, a field that has become known as deep learning (DL) [3–5]. As deep learning surpassed human performance [6, 7], it became more and more popular and was applied in a variety of fields yielding outstanding results in different machine learning applications [8], including speech recognition [9–12], computer vision [13–15], and natural language processing [16–18].

During my Ph.D., I worked closely with the industry partner of my program, Xperi, in solving real problems in consumer technologies using deep learning solutions. Through my collaboration with Xperi and being exposed to industry problems some practical realisation of deep neural network (DNN) based solutions became more and more apparent. A first realisation is related to the evaluation of a model and its use as an industry solution. Nowadays models have achieved state-of-art (SoA) results in different applications [8], but these models are evaluated on a well defined publicly available corpus of data [19]. In industry, one faces unique practical problems such as determining whether one has enough data for a given model, with high quality, labelled data being very expensive to acquire [19]. The methods that are developed and evaluated on the publicly available dataset, no matter how large or

diverse or interesting these datasets are, are clearly not going to be well suited to address the variety of real-life scenarios [19]. For example, an ideal dataset would cover all possible scenarios, but as it is impractical to gather such dataset the reality is that any training dataset provides a finite set of samples of a much larger data distribution. As a result the models are not able to generalise and be robust to different scenarios. Furthermore in industry solutions, lower complexity ANN are preferred to allow for embedded/edge implementation.

The use of data augmentation techniques can provide remedy to some of these problems and in many cases using the correct augmentation techniques can be more important than the selection of the model's architecture [5, 20]. Using augmentation techniques, the dataset size and its variety can be increased [21]. Thus, enhancing the performance of the trained model and also its ability to generalize and be more robust to different scenarios that were not included in the original dataset [21, 22]. Some examples of basic augmentation techniques that are used in most training pipelines of deep learning include adding Gaussian noise, flip, rotate, blur, resize, cutout, crop, zoom in and out. These basic augmentation techniques can be used in almost any problem and in most of the cases it will bring positive results [23].

Another realisation is that when deep learning solutions are applied in industry to solve practical problems, there is a need for custom datasets that are tailored for a particular use case. An example is that image data varies across cameras and thus, algorithms that work on visible images will not work on Near-Infrared (NIR) images, as they are from different domains. In such case a whole new dataset would need to be acquired, which is an engineering nightmare. Additionally, in these cases using the basic augmentation techniques is not always enough to solve the problem but there may be aspects of the practical problems that can be amenable with the use of more advanced data augmentations. For the advanced augmentation techniques, usually an expert who knows the nature of the specific problem in hand designs these to transform the data that are available in order for them to mimic the nature of the specific use case that is being solved [24]. A practical example from [25], where after extensive study of the low-quality images of the customer level iris acquisition handheld devices, a specialised data augmentation pipeline is designed that transforms high-quality iris samples into low-quality ones. Using this augmentation pipeline [25], the author is able to train a CNN that reports the best segmentation accuracy for low-quality iris datasets despite having initially only high quality iris samples.

Practical solutions require appropriate training datasets modified to a constrained use case together with matching ground truth data. Acquiring such datasets at the scale required

for training with today's artificial intelligence (AI) systems and subsequently annotating them with an accurate ground truth is challenging in terms of time, human resources and operational costs. Finally, with the recent introduction of the General Data Protection Regulations (GDPR) and associated complications introduced, I have experienced public datasets being withdrawn [26] and it is getting more and more difficult to gather data that is linked to individual persons. Due to the current difficulties associated with capturing real data, it has become important to find ways to either adapt the data or be able to generate it in controlled ways.

Augmentation techniques can alleviate some of the drawbacks of real samples and assist in the training process, but the fact remains that the augmented samples are still constrained on the data that we already have. Synthetic data gives us the ability to generate new, novel data that do not exist at all in the original dataset. Synthetic data is information that computer simulations or algorithms generate [27]. The recent advancement in the area of synthetic data generation [28, 29], indicate that synthetic data can be used to overcome the disadvantages of real-world data and the challenges related to the data acquisition process. The generation of synthetic data is far cheaper compared to data acquisition and in many cases these data samples come with annotated labels. Furthermore, the computer simulations and/or algorithms that generate synthetic data, can be customised to generate samples for a particular use case, include edge case scenarios and more variation than real-world datasets. Finally, since the synthetic data is generated, there are no underlying GDPR/privacy concerns and they can be used freely. Researchers have been increasingly using synthetic data to train their DNN models [30] and have demonstrated that it can be as good or even better for training an AI model than data based on actual objects, events or people [31]. Furthermore, frameworks and methodologies need to be developed as swapping the real-world data with synthetic is not straightforward. These frameworks have to ensure that the synthetic data behave as the real-world data, mathematically or statistically and that are effective when incorporated in the training pipelines.

1.2   Summary of Contributions in this Thesis

In this section a brief summary of the contributions in this thesis are presented. In the remaining chapters of this thesis the works related to each contribution are thoroughly presented. In each chapter an introductory paragraph provides the context of the research work. Following that, the research objectives/ questions of the work are given, followed by the contributions of the presented research work. Finally a discussion of the contributions is

given, analysing the impact of the contributions to the overall research area. Additionally, for each work, a table is presented, in which the contributions of its author are given related to the four major criteria as explained in section 1.4.

1.2.1 Contribution to Advanced Augmentation Techniques

Chapter 2 contributes to the research area of Advanced Augmentation Techniques. In this research work [32], advanced augmentation techniques are proposed to overcome the unavailability of data for the problem of off-axis iris segmentation in augmented/virtual reality (AR/VR) devices. The proposed augmentation techniques transform high-quality frontal iris images into off-axis of low-quality, to mimic the shape and quality of the iris images when captured from a user-facing camera on an AR/VR device along with their ground truth segmentation map.

Utilising the augmentation techniques, a low-complexity deep neural network is trained, which achieves SoA levels of accuracy in iris region segmentation for the challenging augmented off-axis eye-patches. A network architecture of a lower complexity, compared to the ones proposed in literature, is selected to accommodate the implementation of such algorithm in edge devices. The proposed technique is designed for segmenting off-axis consumer level iris images. Despite that, experiments are carried out on frontal iris images in order to conduct a fair comparison with the other methods trained on such iris images. The proposed method is shown to achieve high levels of performance for regular, frontal, segmentation of iris regions, comparing favourably with SoA techniques of significantly higher complexity. Preliminary results of this work were initially presented in [33].

1.2.2 Contribution to Validation Methodologies for Synthetic Data

The AI tools, based on generative adversarial networks (GANs), for generating synthetic facial samples have evolved [28, 29] in the last years, enabling photo-realistic, high-resolution random synthetic face samples to be generated at scale, with StyleGAN [29] being a representative of the current state-of-art. This leads us to consider the potential to create a large facial dataset built entirely from synthetic facial data samples. In such a case the identity feature is key component and the starting point for building such a dataset is a methodology to demonstrate that the identities of the synthetic facial data samples 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 some key research questions:

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

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

- 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?

These research questions are answered through a series of research works presented in Chapter 3, focused on a methodology validating synthetic data samples.

Initially, in the research work [34], referred to as GADAFAI hypothesis, a discussion is presented on replacing real-world data with synthetic data samples and a roadmap is provided on the steps required to accomplish that. One of main steps on replacing real-world data with synthetic, is to create methodologies and techniques to validate that the synthetic data behave like real ones, statistically or mathematically. In the research paper [35], a methodology and metrics are introduced that allow to validate synthetic data samples in the context of facial biometrics. More specifically, this work explores the identity attribute of synthetic face samples derived from GANs. The methodology, utilising SoA face recognition (FR) algorithm that measures identity similarity, allows 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. Furthermore, using this methodology a technique is provided to remove the most connected data samples within a large synthetic dataset, thus the remaining synthetic face samples can be considered as unique as data samples gathered from different real individuals, in terms of their identity. In Chapter 3, two more research works are presented [36, 37], which are initial building blocks for creating the validation methodology and metrics for [35].

1.2.3   Utilising Synthetic Data in Training Pipelines for Edge-AI Solution

In Chapter 4, the effective use of advanced AI tools is shown, in order to overcome the unavailability of real-world samples and allow to study the investigate the task in hand and train effectively a convolutional neural network (CNN). More specifically, in [38], the lack of real face samples with enough illumination variation is observed. To overcome this obstacle a SoA re-lighting technique is utilised to augment face samples and introduced different illumination scenarios. This allows the study of the effect of each illumination

scenario on the performance of a SoA FR algorithm. Furthermore, by utilising the re-lighting technique there are sufficient face samples with several illumination scenarios allowing to explore the possibility of fine-tuning the FR model, in order to be robust in such conditions. The experiments showed that the fine-tuned FR model can cope with different illumination variations when trained with sufficient data that represent the problem correctly.

Furthermore, currently I am collaborating/ mentoring a Ph.D. student and additional research papers are in preparation, which are the continuing of this work [38], presented in Chapter 4.

1.2.4   Contributions to Training Methodologies for Conditional GANs

In Chapter 5, a new framework is presented to train a deep conditional generator by placing a classifier or regression model in parallel with the discriminator and back propagate the classification or regression error through the generator network [39]. Special cases for binary classification, multi-class classification, and regression are studied. Experimental results on several datasets are provided and the results are compared with similar SoA techniques. The main advantage of the method is that it is versatile and applicable to any variation of GAN implementation but also it is shown to obtain superior results compared to other methods. The mathematical proofs for the proposed schemes for both classification and regression are presented.

1.2.5   Other Contributions

In Chapter 6, some of secondary publications are presented. *Deep Learning for Consumer Devices and Services* is a series of articles in which the main focus of the discussion is on the deployment of deep learning on consumer devices. This series has currently four articles, from which I was involved in two of them [40, 41].

In [40, 41], a discussion is presented regarding the shift of AI from the data center into consumer electronics (CE) devices—"Edge-AI" and some of the challenges that data acquisitions entail, are presented. Furthermore an initial introduction of basic and advanced augmentation techniques is given, along with their benefits.

1.3   List of Publications

Contribution to Advanced Augmentation Techniques

1. Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets." Neural Networks 121 (2020): 101-121

2. Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "A deep learning approach to segmentation of distorted iris regions in head-mounted displays." In 2018 IEEE Games, Entertainment, Media Conference (GEM), pp. 1-9. IEEE, 2018

Contribution to Validation Methodologies for Synthetic Data

3. Corcoran, Peter, Hossein Javidnia, Joseph E.Lemley, and Viktor Varkarakis. "Generative Augmented Dataset and Annotation Frameworks for Artificial Intelligence (GADAFAI)." In 2020 31st Irish Signals and Systems Conference (ISSC), pp. 1-6. IEEE, 2020

4. Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "Re-Training StyleGAN-A First Step Towards Building Large, Scalable Synthetic Facial Datasets." In 2020 31st Irish Signals and Systems Conference (ISSC), pp. 1-6. IEEE, 2020

5. Varkarakis, Viktor, and Peter Corcoran. "Dataset Cleaning—A Cross Validation Methodology for Large Facial Datasets using Face Recognition." In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6. IEEE, 2020

6. Varkarakis, Viktor, Shabab Bazrafkan, Gabriel Costache, and Peter Corcoran. "Validating seed data samples for synthetic identities–methodology and uniqueness metrics." IEEE Access 8 (2020): 152532-152550

Utilising Synthetic Data in Training Pipelines for Edge-AI Solution

7. Varkarakis, Viktor, Wang Yao, and Peter Corcoran. "Towards End-to-End Neural Face Authentication in the Wild–Quantifying and Compensating for Directional Lighting Effects." arXiv preprint arXiv:2104.03854 (2021)

Contributions to Training Methodologies for Conditional GANs

8. Bazrafkan, Shabab, Viktor Varkarakis, Joseph Lemley, Hossein Javidnia, and Peter Corcoran. "Versatile Auxiliary Classification and Regression With Generative Adversarial Networks." IEEE Access 9 (2021): 38810-38825.

Other Contributions

9. Corcoran, Peter, Joseph Lemley, Claudia Costache, and Viktor Varkarakis. "Deep Learning for Consumer Devices and Services 2—AI Gets Embedded at the Edge." IEEE Consumer Electronics Magazine 8, no. 5 (2019): 10-19

10. Corcoran, Peter, Claudia Costache, Viktor Varkarakis, and Joseph Lemley. "Deep learning for consumer devices and services 3—Getting more from your datasets with data augmentation." IEEE Consumer Electronics Magazine 9, no. 3 (2020): 48-54

1.4   Contribution Taxonomy

As this publication based thesis includes work that was done in collaboration with others, this section provides an overview of main criteria that determine primary authorship. The CRediT approach [42] has been adopted by journals in several fields to specify the contributions of individual authors. In the CRediT Taxonomy each author's contributions are measured as a percentage point on 14 roles. These are: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

In this thesis, despite the collaborations, the majority of the work is my own, therefore a more concise generalization of this taxonomy which encapsulates the major criteria is adopted similar to [43]. More specifically:

1. Research Hypothesis/ Idea.

2. Methodology, which includes validation, data curation, formal analysis, tool selection, software development, implementation and experiments.

3. Background which includes, investigation, formalization and work done to place the research efforts in a wider context of literature in a given field, this may include some aspects of writing (literature reviews) and informs aspects of project administration and supervision and ensuring that methodology used is typical of that used in the area of publication.

4. Manuscript Preparation which includes all aspects of writing manuscript preparation including Writing – original draft, Writing – review & editing, and Visualization except those specified in the next criteria.

This generalization has the weakness that it ignores most aspects of funding, project administration, resources or supervision but otherwise encapsulates the main points that would determine primary authorship. Such a table will be presented in each main work presented in this Thesis, attributing the contribution of each author to the aforementioned four criteria. Authors are listed by initial where VV means Viktor Varkarakis, SB means Shabab Bazrafkan, PC means Peter Corcoran, HJ means Hossein Javidnia, GC means Gabriel Costache and WY means Wang Yao. Contribution percent is listed at a resolution of %.

Chapter 2

Contribution to Advanced Augmentation Techniques

In this chapter an overview of the research work: *Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets." Neural Networks 121 (2020): 101-121*, [32], is given along with its research objectives and a discussion on the primary contributions. The copy of the paper published based on this section is presented in the Appendix A. Preliminary results of this work were initially presented in the: *Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "A deep learning approach to segmentation of distorted iris regions in head-mounted displays." In 2018 IEEE Games, Entertainment, Media Conference (GEM), pp. 1-9. IEEE, 2018*, [33], which can be found in the Appendix B.

The contributions of its author related to the four major criteria as explained in section 1.4, for the research works [32, 33] are presented in Table 2.1

Table 2.1 Authors' Contributions to [32, 33]

Contribution Criteria	Contribution Percent
Research Hypothesis/Idea	VV 70%, SB 30%
Experiments and Implementation	VV 80%, SB 20%
Background	VV 70%, SB 30%
Manuscript Preparation	VV 70%, SB 15%, PC 15%

Nowadays, the authentication requirements in consumer devices are evolving beyond today's mobile devices. New AR/VR headsets provide a gateway to sophisticated virtual worlds and online services [44–46]. A key challenge with AR/VR headsets is that they do not provide an intuitive mean of user authentication [47, 48]. User-facing camera systems

Fig. 2.1 AR/VR devices with a user-facing camera.

can be incorporated into such headsets as shown in Fig. 2.1, thus making feasible the implementation of an iris authentication mechanism. In such scenario, the iris images obtained are characterised as off-axis. The current segmentation solutions, used in today's iris authentication pipelines, are optimised for frontal iris images which rarely appear in the AR/VR scenarios. Given the off-axis shape of the iris images in such scenarios, an accurate segmentation would be challenging. Furthermore in the iris authentication workflow failed segmentations represent the single largest source of error [25, 49–52]. Therefore, it is identified that a new iris segmentation solution for the AR/VR cases needs to be developed in order for iris authentication to be effective in wearable headsets.

2.1   Research Objectives

The iris segmentation method developed for AR/VR cases needs to be of high performance and robust with off-axis iris images. A dataset with iris images captured from the front-facing camera of these devices along with their corresponding segmentation ground truth is not available. As a result, the main obstacle in developing an iris segmentation solution for the AR/VR cases is the unavailability of data representing the problem correctly. Acquiring such a dataset is challenging due to GDPR but also due to the manual effort needed to annotate/label the ground truth segmentation maps. As there are several publicly available frontal iris datasets with ground truth [53–55], the use of augmentation techniques can be utilised to overcome the unavailability of data and produce a large dataset that represents the problem correctly.

Furthermore, existing segmentation solutions consist of networks with high complexity and memory requirements, making it difficult for deployment in edge devices such as AR/VR headsets, which was initially discussed in [25]. As a continuation of the work in [25], we investigate the potential of choosing a network architecture with reduced complexity, to facilitate the implementation of such solutions in AR/VR headsets, while being able to preserve the performance and be competitive when compared to larger CNN architectures.

2.2   Summary of Contributions

In order to accomplish the research objectives, the first step is to overcome the unavailability of data by generating a large number of samples, representing off-axis iris images as captured by a user-facing camera from an AR/VR device. Now, frontal iris images, are usually captured in a controlled environment [53, 54] and positioned in the centre of the image, having a circular shape. After extensive study of iris images from AR/VR cases, the main differences compared to the common frontal iris images in terms of shape are as follows:

- such iris images are off-axis in the horizontal and vertical plane

- characterised with a distorted, elliptical shape and

- not in the centre of the image (as usual).

In our solution we propose:

1. Two new specialised augmentation techniques, that take as an input high-quality frontal iris images and produce an iris image which simulates the characteristics of iris images as represented when captured from a user-facing camera on an AR/VR device along with their ground truth segmentation map (Appendix A, section 2.3.2). In addition, augmentation techniques that transform high-quality iris images into diverse low-quality iris images are borrowed from [25] to represent unconstrained acquisition conditions that occur from images captured from a user-facing camera on wearable AR/VR devices (Appendix A, section 2.3.1).

2. A pipeline of combining these augmentation techniques is designed (Appendix A, section 2.3.3), which allows to increase the number of samples available and result in a dataset that represents a generalized and realistic scenario of iris images from AR/VR devices in order to train a CNN model effectively.

3. Finally, an improved low complexity neural network design for the off-axis iris segmentation task is proposed. The proposed network is targeted for deployment on embedded wearable headsets, as it has reduced memory and computational requirements in comparison with other deep learning SoA iris segmentation techniques (Appendix A, section 3).

The proposed network has the best performance on segmenting the augmented off-axis iris samples (Appendix A, section 5.1). Interestingly, the segmentation performance of this network on frontal iris samples from several public datasets, is comparable with the

current SoA iris segmentation SPDNN network [25] (Appendix A, section 5.2), despite its complexity being at least an order of magnitude less than the SPDNN [25] (Appendix A, section 3.2).

The code and instruction on how to use the proposed augmentation techniques along with the weights of the trained network can be found in the GitHub repository of this research work [1].

2.3   Discussion of Contributions

This work provides an iris segmentation solution specialised for the new AR/VR devices. Through this work it is clear that in order to provide a suitable solution, the main bottleneck as in most machine learning (ML) solutions is data availability. Utilizing though the current available data and the new specialised augmentation techniques proposed it is shown that is possible to overcome the unavailability of data for a different problem of the same domain and avoid the challenging procedure of the data acquisition and labelling. Despite the fact that in this work the augmentations techniques are focused on generating off-axis iris images, they could be widely applicable for other off-axis region problems or/and cases where the images are stretched (e.g.: when using cameras with a wide field of view).

Also, nowadays it is common practice that when an architecture of CNN is designed, that deep and large structures are favoured. This is preferred in order to increase the possibility of solving the investigated problem or promise higher performance from a smaller size CNN. Selecting a CNN with a deeper structure rather than a more compact structure, comes with drawbacks such as increased training and execution time as well as generous memory requirements, which are not always available in edge devices. This work showcases how incorporating the correct data augmentation techniques in the deep learning solutions can be more important than having a complex model [5, 20] and help design solutions suitable for edge devices.

This is illustrated through the comparisons between our proposed network and the SoA solution in the task of segmenting frontal iris samples, SPDNN [25]. The SPDNN [25] has a complex network architecture with more than 1M parameters while the proposed network consists only of 70k parameters (Appendix A, section 3.2). The proposed network, despite having $\times 10$ less parameters and focused on segmenting off-axis iris images, its performance is comparable with the SPDNN [25] in several frontal iris datasets (Appendix A, section 5.2.2). The main difference between them except their complexity, is that the

---

[1]Code available at: https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Off_axis_Iris

proposed network is trained with frontal and off-axis iris images while the SPDNN [25] is only trained with frontal iris images. The proposed augmentations techniques (that transform the frontal iris samples to off-axis, Appendix A, section 2.3.2), not only increased the number of samples available to train the network compared to the number of samples used to train SPDNN [25], but also injected variety and randomness to the data samples that the samples used to train SPDNN [25] lack. Utilising these components, it allowed us to train a lower complexity network while keeping the performance comparable to the higher complexity, SoA SPDNN [25]. Thus, showing the importance of appropriate data augmentation strategies over heavy-weight network structures, that allow to design solutions that can be efficiently installed in edge devices, without losing significant performance.

Chapter 3

Contribution to Validation Methodologies for Synthetic Data

3.1 Introduction

Synthetic Data and Tools

Synthetic data is information that computer simulations or algorithms generate as an alternative to real-world data [27]. In this Chapter, we focus on the synthetic data generated by algorithms.

Some of the most popular tools for synthetic data generation, are known as GANs. GANs [56] utilise Deep Neural Network capabilities and are able to estimate the data distribution for large size problems. These models comprise two networks, a generator, and a discriminator. The generator makes random samples from a latent space, and the discriminator determines whether the sample is adversarial, made by the generator, or is genuine image coming from the dataset [24]. The general framework of GANs is shown in Fig 3.1.

Initially, despite the potential of GAN models, their use remained mostly within the research community due to some of their drawbacks [58, 59]. In general, GANs are known for their difficulty to train [58, 59], but the main reason they did not become directly accepted by the industry for commercial use, was due to their output. In the early days of GANs the output images were of low quality, small resolution and more importantly unrealistic [56, 60]. When observing an image generated from a GAN, it was easily recognisable that these images were fake [56, 60]. Over time, GANs evolved and the current SoA GANs are able to generate high-quality, realistic samples in a variety of image resolutions which can be used effectively in a range of applications [28, 29]. Progressive-GAN [28] and StyleGAN
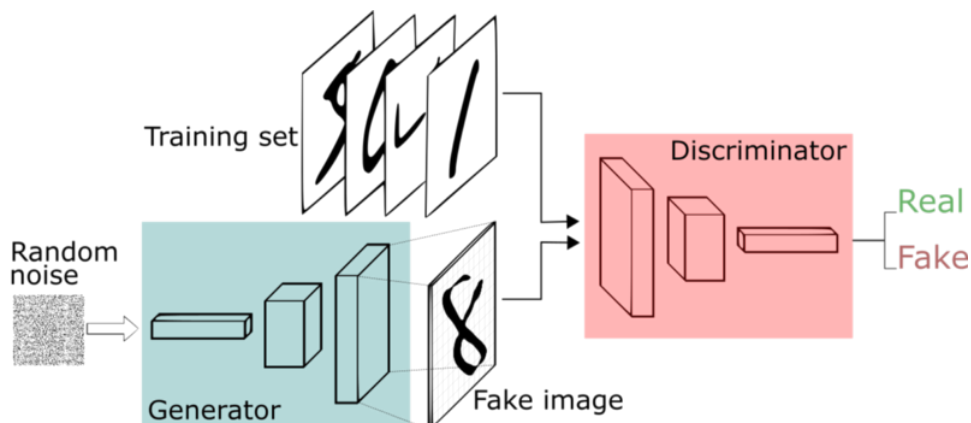
Fig. 3.1 Generative Adversarial Network framework [57].

[29], were one of the first methods that with high-quality, realistic outputs and became the de facto benchmarks for GANs. In Fig. 3.2, the evolution of GANs in the task of face generation is shown, going from low resolution, blurry and unrealistic face samples to face images that can trick the human perception in to believing that these samples are real [61].

Fig. 3.2 The progress of GANs in the task of face generation [62].

The Challenges of 'Real-World' Data

It is known that data acquisition, availability and preparation are the biggest bottlenecks in the ML/DL pipelines [63], which was also shown through the research work described in Chapter 2. An ideal dataset would cover all the likely sensing and environmental contexts that might arise, but as it is impractical to gather data for all the possible cases, real-world datasets are sparser compared to an ideal scenario. As an extension (as the data do not cover all the possible scenarios) bias is introduced to these datasets which is then inherited by

our trained algorithms. Furthermore, due to the fast-changing technology environment, new unique problems arise daily. In order to realise solutions in many of these specific problem domains there is a growing need to build custom datasets that are tailored for a particular use case. Practical solutions require appropriate training datasets modified to a constrained use case together with matching ground truth data. Acquiring such datasets at the scale required for training with today's AI systems and subsequently annotating them with an accurate ground truth is challenging in terms of time, human resources and operational costs. Finally, with the recent introduction of GDPR and associated complications introduced, industry now faces additional challenges in the collection of training data that is linked to individual persons.

Why Synthetic Data

The improvements in synthetic data generation in recent years suggest that synthetic data can now be used to overcome obstacles and disadvantages of real-world data. The generation of synthetic data is far cheaper compared to data acquisition and in many cases these data samples come with annotated labels. Furthermore, the synthetic models can be modified and customised to generate samples for a particular use case, include edge case scenarios and more variation than real-world datasets. Finally, but equally importantly, since the data is generated, there are no underlying GDPR/privacy concerns and they can be used freely.

As a result the use of synthetic data, is a promising technique on the rise in modern deep learning. Researchers have been increasingly using them to train their DNN models [30] and have demonstrated that it can be as good or even better for training an AI model than data based on actual objects, events or people [31]. Furthermore, frameworks and methodologies needs to be developed as swapping the real-world data with synthetic is not straightforward. These frameworks have to ensure that the synthetic data behave as the real-world data, mathematically or statistically and that are effective when incorporated in the training pipelines.

Research Questions

The development in the area of synthetic data and the real-world data related difficulties that research and industry community is facing, suggest that it might make sense to focus on developing improved methodologies to control and manage the generation of data samples matched to a specific machine learning problem rather than struggle with the challenges of obtaining sufficient 'real-world' data. These ideas lead to the following research questions:

- Can we artificially generate and/or augment suitably large sets of data samples adapted for training today's AI networks?

- Can we prove that the resulting AI networks are as robust and reliable as those trained on equivalent 'real-world' datasets?

These research questions are introduced and analyzed in the work presented in section 3.2, referred to as GADAFAI hypothesis [34]. In that work, a plan is presented with the steps required in order to validate the hypothesis and incorporate effectively the AI tools for data generation in the training pipelines and replace real data. The first steps can be divided in two categories:

- Building AI tools, that will generate a large amount of new synthetic data and/or augment these data samples.

- Create methodologies and techniques to validate that the synthetic data behave like real ones and also measure whether their use is effective when incorporated in the training pipelines.

Some practical examples and preliminary results are given in the work presented in section 3.2 in the context of facial biometrics, with the ultimate goal of creating a large synthetic face dataset.

In the remainder of this chapter (sections 3.3-3.5), the research work presented contributes on the methodology side of the synthetic data, as techniques are proposed that aim to validate whether or not the synthetic data "behave" like real-world data.

The methodology developed is in the context of facial biometrics as it is one of best starting points, where synthetic might replace real data samples. That's because facial biometrics is a topic that has been heavily researched in the literature and there is a wide variety of large face datasets that are easily accessible. Furthermore the AI tools related to the facial biometrics such as Face Recognition algorithms or GANs [29] for the task of face generation have improved significantly compared to other areas. Finally, but equally important is the fact that facial samples is one of the most susceptible data categories to GDPR and privacy regulations.

The formulation of different building blocks necessary to create the aforementioned validation methodology, are given below. More specifically, in section 3.3, the expansion of AI tools for facial data generation is presented. In the section 3.4, irregularities are found and corrected in real-world datasets. This allows for a clear/correct distribution of real data, thus when creating methodologies to validate the synthetic data, fair comparisons can be made

between the real and synthetic data. In section 3.5 , by incorporating parts from the works of 3.3 and 3.4 new methodologies and metrics are proposed that validate the synthetic data.

In each section the research objectives/questions and summary of contributions for each individual work are given. Finally, in section 3.6, a discussion is presented, which focuses on the how the individual sections relate to one another and their overall contribution as a unified research work.

3.2 Framework for Use of Synthetic Data and Data Augmentation Techniques for Improved Training of AI Networks

In this section an overview of the research work: *Corcoran, Peter, Hossein Javidnia, Joseph E. Lemley, and Viktor Varkarakis. "Generative Augmented Dataset and Annotation Frameworks for Artificial Intelligence (GADAFAI)." In 2020 31st Irish Signals and Systems Conference (ISSC), pp. 1-6. IEEE, 2020.*,[34], is given along with its research questions and a discussion of contributions. The copy of the paper published based on this section is presented in the Appendix C. The contributions of its author related to the four major criteria as explained in in section 1.4, for this research work [34] are presented in Table 3.1.

Table 3.1 Authors' Contributions to [34]

Contribution Criteria	*Contribution Percent*
Research Hypothesis/Idea	PC 70%, HJ 30%
Experiments and Implementation	VV 100%
Background	PC 60%, HJ 40%
Manuscript Preparation	PC 60%, HJ 20%, JL 10%, VV 10%

3.2.1 Research Questions and Objectives

Due to the increasing challenges in the collection of real-data, the idea of using advanced AI tools, such as GANs and augmentation techniques, to generate new data tailored for training machine learning algorithms and eliminate the need for real-world data, leads to somewhat contrarian research questions (hereafter referred to as the "GADAFAI hypothesis") (Appendix C, section I-A):

- Can we artificially generate and/or augment suitably large sets of data samples adapted for training today's AI networks?

- Can we prove that the resulting AI networks are as robust and reliable as those trained on equivalent 'real-world' datasets?

One of the main goals of GADAFAI [34], is to build synthetic datasets, but is distinguished from other similar efforts in that it also seeks to provide some validation metrics to measure the usefulness of these synthetic data samples for a particular use case or problem. These metrics should allow researchers to quantify variances between 'real-world' datasets and those that are 'generated' via alternative methodological approaches. The intention is to focus on measuring the validity of synthetic datasets for practical problems across a range of fields of application. This should enable new methodological refinements of the generated datasets to specific use cases.

3.2.2   Summary of Contributions

In the work presented in [34], the challenges in acquiring large datasets of 'real-world' data are discussed (Appendix C, section II) along with a high level review of current SoA in data generation and augmentation techniques, which are the foundation for the GADAFAI framework in order to build large synthetic datasets (Appendix C, section III). This is followed by a discussion of the primary research domains and associated data landscapes in the context of today's computer vision research, in which the GADAFAI hypothesis [34] can be useful and the challenges associated with each domain. Furthermore, how GADAFAI [34] can work in practice is discussed, providing a roadmap towards a broader validation of the hypothesis and answering the research questions posed (Appendix C, section IV). Through preliminary results an initial approach and the steps required to validate the GADAFAI hypothesis [34] in the context of facial biometric data are outlined (Appendix C, section V).

3.3   Understanding the AI Tools for the Task of Face Generation - Re-Training StyleGAN

In this section an overview of the research work: *Varkarakis, Viktor, Shabab Bazrafkan, and Peter Corcoran. "Re-Training StyleGAN-A First Step Towards Building Large, Scalable Synthetic Facial Datasets." In 2020 31st Irish Signals and Systems Conference (ISSC), pp. 1-6. IEEE, 2020.*, [36], is given along with its research objectives and a discussion of the

contributions. The copy of the paper published based on this section is presented in the Appendix D. The contributions of its author related to the four major criteria as explained in in section 1.4, for this research work [36] are presented in Table 3.2

Table 3.2 Authors' Contributions to [36]

Contribution Criteria	*Contribution Percent*
Research Hypothesis/Idea	VV 80%, SB 10%, PC 10%
Experiments and Implementation	VV 90%, SB 10%
Background	VV 90%, SB 10%
Manuscript Preparation	VV 70%, SB 15%, PC 15%

The original GAN presented in [56] is made of two Deep Neural Networks: a generator and a discriminator. The generator accepts a tensor of randomly generated numbers and returns an image and the discriminator is a binary classifier that accepts an image and determines whether it is a generated image or not. In this approach, these two networks are trained in a min-max game wherein the final goal is for the generator to synthesis an image that the discriminator classifies as a real image. StyleGAN [29] is one of the variations of GAN wherein the generator is developed in a specific way which separates it from its preceding implementations in three main ways:

- The latent space (Z) is reshaped via a fully connected DNN (which returns W) before feeding into the generator. This is to introduce disentanglement to the original latent space (Z) during the mapping into style indicators (W) [29].

- The latent space is not fed into the generator at its input layer. The new latent space W is given to the generator before each convolutional layer. In other words, each part of the vector W is induced into the generator in a different layer. This gives the opportunity to introduce style information at different levels [29].

- A Gaussian noise is added to the features before each convolution. This operation helps the network to use its maximum capacity and generate higher quality outputs with high-frequency features [29].

This modifications and novelties in the design of GANs have contributed into making StyleGAN [29] one of the most sophisticated and successful AI tools. StyleGAN [29] is able to generate high quality and resolutions samples and has been trained on different image topics such as cats, cars, bedrooms. Although, the models that really caught the eye of the public, are the ones trained on the face datasets of FFHQ [29] and CelebA-HQ [28]. These

models are able to generate high-quality, high-resolution, but also realistic face samples that are able to fool humans in many cases [61].

StyleGAN [29] is the perfect example of advanced AI tools that can be used in generating synthetic data and the GADAFAI [34] framework. As discussed in 3.1, we seek to validate the synthetic data in the context of facial biometrics with the ultimate goal of generating a large synthetic face dataset and the official models of StyleGAN [29] trained on the face datasets of FFHQ [29] and CelebA-HQ [28], provide a good starting point. Despite that, more models need to be trained with larger datasets in order to understand StyleGAN [29] better, in order to use effectively its samples in the data pipelines.

3.3.1 Research Objectives

As a first step, in the process of validating the GADAFAI hypothesis [34], it is necessary to build and understand the tools that can generate a larger number of unique synthetic data samples of a human face. Using StyleGAN [29] to generate new face samples we need to understand its capabilities and drawbacks. At the time of the research, StyleGAN [29] had been recently released making available two models, trained on FFHQ [29] and CelebA-HQ [28], which are datasets with 70k and 30k samples. However, these are not particularly large facial datasets with relatively low number of identities. Therefore, in order to explore StyleGAN [29], there was a need for more StyleGAN models trained on several datasets with more samples, identities and different image resolution and quality, which inspired the research work of this section. In that way, we can investigate how the quality, attributes, variety and other characteristics of the generated face samples are related to the datasets used. Furthermore, having more StyleGAN models [29] trained in a variety of datasets quality and size, will allow the general validation of the GADAFAI hypothesis [34] in context of facial biometrics and it wont be constrained only to the initial StyleGAN models [29].

3.3.2 Summary of Contributions

In this research work [36], the procedure of re-training StyleGAN [29] on two publicly available datasets, CelebA [64] and CASIA-WebFace [65], is presented. CASIA-WebFace [65] contains almost 500k face samples from more than 10k identities, while CelebA [64] consist of 200k face samples from 10k identities. These datasets are larger compared to the FFHQ [29] and CelebA-HQ [28] but also of lower resolution and quality, thus the StyleGAN models of this work [36] are trained on a 256x256 image resolution. Furthermore, practical issues and challenges arising from the retraining process are discussed (Appendix D, section III). Tests and validation results are presented and a comparative analysis of the re-trained

StyleGAN models is provided (Appendix D, section IV). Finally, the resulted StyleGAN
models are made publicly available and can be found at [2]. The official StyleGAN models
and implementation code can be found at [3].

This work [36], expands the available StyleGAN models [29] for generating face samples.
This is one of the first building blocks in order to validate synthetic samples and part of
the GADAFAI hypothesis [34] in the context of facial biometrics. Furthermore, StyleGAN
models [29] are expensive computationally but also in time, (e.g.: a machine with a single
V100s GPU, will need more than two week of computational time for a lower resolution
(256x256) StyleGAN model), therefore the trained models are made publicly available in
order to facilitate and support research in the domain.

3.4 Correcting the Distribution of Real World Samples - A Methodology for Cleaning Large Facial Datasets

In this section an overview of the research work: *Varkarakis, Viktor, and Peter Corcoran.
"Dataset Cleaning—A Cross Validation Methodology for Large Facial Datasets using Face
Recognition." In 2020 Twelfth International Conference on Quality of Multimedia Experience
(QoMEX), pp. 1-6. IEEE, 2020.*, [37], is given along with its research objectives and
a discussion of contributions. The copy of the paper published based on this section is
presented in the Appendix E. The contributions of its author related to the four major criteria
as explained in in section 1.4, for this research work [37] are presented in Table 3.3.

Table 3.3 Authors' Contributions to [37]

Contribution Criteria	*Contribution Percent*
Research Hypothesis/ Idea	VV 90%, PC 10%
Experiments and Implementation	VV 100%
Background	VV 100%
Manuscript Preparation	VV 80%, PC 20%

---

[2]Code available at: https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Re-training-StyleGAN
[3]Code available at: https://github.com/NVlabs/stylegan

3.4.1 Research Objectives

In recent years, large "in the wild" face datasets have been released in an attempt to facilitate progress in tasks such as face detection, face recognition, and other tasks [66–68]. Generally, large face datasets are built in a semi-supervised way using image-search engines and thus prone to bad data samples due to mislabelling and poor image quality of some samples [69–71]. Often, the number of these bad data samples is not statistically significant for a particular task and they can be ignored, but, in other cases a small number of bad data samples can become quite significant and lead to poor training outcomes [72].

Consider for example a large dataset which is used to train a face detector, if the training dataset contains some mislabelled identities – i.e. wrong identity is assigned to a person – this is not critical for training a face detector, as these mislabelled data samples still represent 'good' samples of facial images, as their identity characteristic is not used. However, if the task at hand switches to training a CNN to perform facial recognition, distinguishing between multiple identities, these mislabelled samples are now 'bad' and if there are sufficient mislabelled data samples, the performance of the resulting face recognition CNN will be sub-optimal [72]. Therefore, cleaning the datasets from mislabelled samples is desirable for some use cases.

3.4.2 Summary of Contributions

In this work [37] a (semi-automatic) technique for identifying and removing mislabeled samples in terms of identity is developed. The technique utilizes a face recognition model trained / fine-tuned on the examined dataset in order to discover outliers in an identity folder that shall be examined as it is possible to contain mislabeled face samples (Appendix E, section III). This methodology is applied to clean the CelebA dataset [64] and show its effectiveness (Appendix E, section IV). This technique can be applied to any face dataset annotated with identities in order to "clean" it so that the dataset can be used with more certainty as a considerable number of mislabeled samples will be eliminated and as a result the training algorithms and validation metrics are more accurate.

Similar methods exist [72–75], although these can clean a large part of the datasets, they have some drawbacks as after the cleaning procedure, the datasets may either lack diversity as many variations are treated as outliers or the size of the dataset has decreased quite significantly due to the rigorous constraints imposed by the cleaning process. This research work [37], is distinguished from other methods in that we seek to provide a minimal curation of the dataset, in order to retain as many original data samples as possible to ensure that the diversity of the original dataset is preserved.

The list of the mislabelled samples found in the CelebA dataset [64], using the proposed method [37] is made available and can be found at the GitHub repository of this work [4].

3.5 Creating Methodologies and Metrics to Validate Synthetic Data Samples

In this section an overview of the research work: *Varkarakis, Viktor, Shabab Bazrafkan, Gabriel Costache, and Peter Corcoran. "Validating seed data samples for synthetic identities–methodology and uniqueness metrics." IEEE Access 8 (2020): 152532-152550.*, [35], is given along with its research questions, objectives and a discussion of contributions. The copy of the paper published based on this section is presented in the Appendix F. The contributions of its author related to the four major criteria as explained in in section 1.4, for this research work [35] are presented in Table 3.4

Table 3.4 Authors' Contributions to [35]

Contribution Criteria	*Contribution Percent*
Research Hypothesis/ Idea	VV 80%, SB 10%, PC 10%
Experiments and Implementation	VV 80%, SB 10%, GC 10%
Background	VV 60%, SB 20%, GC 20%
Manuscript Preparation	VV 70%, SB 10%, PC 20%

As part of GADAFAI [34], this chapter explores the potential to build synthetic facial datasets at scale by using a GAN to generate a seed dataset of facial data samples. Given such dataset, it would then be feasible to modify these seed samples to build a large synthetic training datasets focusing on other facial attributes such as facial lighting, pose, and expression [76–78]. When considering the real face data samples, one of the most important attributes is their identity. The identity attribute is unique for each subject. In order for the synthetic face dataset to be valid, it is needed to demonstrate that the identity attribute of the synthetic facial data samples used in the seed dataset behave in the same way as those of a 'real-world' face dataset, as it is essential for the synthetic data to reflect the attributes of real data as discussed in 3.1.

---

[4]Code available at: https://github.com/C3Imaging/Deep-Learning-Techniques/tree/clean-celebA

3.5.1   Research Questions

The starting point for building a valid seed dataset using synthetic face samples 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, meaning that the identity attribute is unique for each seed sample. It is also essential to validate that the synthetic data samples are unique in terms of identity when compared with the original seed data used to train the generator, in order to avoid any GDPR/ privacy issues and be able to use them freely. These considerations lead to three key research questions (Appendix F, section I-A):

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?

3.5.2   Summary of Contributions

These research questions led us to develop two approaches to understand and quantify the identity uniqueness of a set of synthetic data samples when compared against samples from the seed dataset. The same methodology can then be applied to understand the uniqueness within a set of synthetic data samples, in terms of identity, when comparing with one another. To evaluate the identity uniqueness of the generated synthetic samples for both cases, the proposed approaches utilize a SoA 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 these research questions (3.5.1).

In the first approach, the performance/behaviour of real samples and generated synthetic samples for each examined case is illustrated through receiver operating characteristic (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 (Appendix F, section IV-B).

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, from a pair of samples, using their embeddings derived from the FR tool used, a score is obtained and it is compared to an FR threshold to determine their identity similarity, which reversely shows the identity uniqueness. As an extension, the second approach enables us to find and quantify the unique samples in a generated synthetic set of samples when compared with either the seed dataset or within itself. This can be used as a metric, to measure the ability of a GAN model to generate synthetic data with unique identities. The FR threshold used to implement this approach is representative of the statistical behaviour of a dataset of real face samples (Appendix F, section IV-C).

The two proposed approaches are implemented using the generated samples from different StyleGAN [29, 36] models, in order to answer the research questions (3.5.1). It should be mentioned that the proposed methodologies can be implemented with any GAN model trained for the task of face generation. The identity uniqueness of the synthetic data samples when compared to the seed data and when compared to one another, is examined . 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, as the synthetic data did not inherit the identity attribute from the seed dataset and can be used without any privacy issues.

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, showing that synthetic samples exist that share the same identity attribute. Finally, the metrics introduced in the second approach (where the thresholding technique is used), that show the ability of the models to generate unique synthetic samples are used. The metrics revealed that in some cases only 7-9% of the samples have a unique identity from a set of 20k generated synthetic samples. Detailed information of the experiments can be found in the Appendix F, section V.

Face samples with unique identity, generated from different StyleGAN [29] models, which can form a seed synthetic face dataset with distinct identities can be found at the GitHub repository of this work [5].

---

[5]Code available at: https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Synthetic_Face_Datasets

3.6 Discussion of Overall Contributions

In this chapter, we introduce the GADAFAI hypothesis [34] and implement some of parts of the required steps in order to validate the proposed hypothesis but also the synthetic data in general.

More specifically, section 3.2, explains and analyses the GADAFAI hypothesis [34], which its main goal is to build synthetic datasets but also provide validation metrics to measure the usefulness of these synthetic data samples in the ML/DL training pipelines. A generalised framework is presented, on the required steps needed, so that synthetic data from today's advanced AI tools can eventually replace the need for 'real-world' data, which is the main bottleneck in today's machine learning era. Ultimately, this work showcases how through creating methodologies to validate the proposed hypothesis, the capabilities of So neural networks can be enhanced and ultimately, such methodological frameworks could free researchers from concerns with the logistics of dataset acquisition, enabling them to focus on new technology innovations in terms of smart services and products.

Following the introduction of the GADAFAI framework [34],in the rest of the chapter the presented works (3.3-3.5) ultimately contribute to the GADAFAI hypothesis [34] and to the research area of synthetic data, as a methodology is proposed that validates whether or not the synthetic data behave like real ones, in the context of facial biometrics with the ultimate goal of creating a large synthetic face dataset.

A first step in order to create the aforementioned validation methodology, is to understand the tools selected to generate a larger number of synthetic face data samples. The main tool selected for that purpose, is StyleGAN [29] as it represents the SoA GAN for the face generation task, although any GAN model trained on this task can be used. As StyleGAN [29] is selected to generate new face samples we need to understand its capabilities and drawbacks. The official StyleGAN models [29] trained on FFHQ [29] and CelebA-HQ [28], are a good starting point, but these datasets are relatively small and the samples are of high quality, which is not usual for face datasets. Therefore there is a need to train StyleGAN models [29] with larger datasets of lower resolution, in order to examine how StyleGAN behaves when trained with such face datasets. The work described in section 3.3, implements that, in which StyleGAN is re-trained on two large face datasets. In that way, we can explore how StyleGAN performs when trained with different datasets. At the same time, this work expands the variety of available models. This allows for the validation methodology to be generalised and not constrained to the initial StyleGAN models [29], that are trained with a datasets that have similar characteristics (relatively small amount of samples and high

quality-resolution).

The validation methodology that we seek to develop, has to determine whether or not the synthetic data "behave" as the real data. This can be investigated in terms of selected characteristic(s) of the samples. In terms of our use case (generating a large synthetic face dataset), when considering the real face data samples, one of the most important attributes is their identity, which is unique for each subject. In order for the synthetic face dataset to be valid, a methodology is needed to demonstrate that the identities of the synthetic facial data samples used in the dataset behave in the same way as those of a 'real-world' face dataset.

A necessary step for forming the validation methodology, is to measure the performance/ behaviour of the real-world face samples in terms of the selected characteristics(s), in our case their identity, and create a reference point/ benchmark which the synthetic data will be compared against it, in order to determine whether or not they behave as the real face samples. The selected tool to develop this methodology and measure the performance/ behaviour of the real face samples in terms of their identity, is a SoA face authentication algorithm. Current face authentication algorithms have evolved and are exceeding the 99% accuracy in most of the datasets [79] thus making it a reliable tool. We started measuring the performance of the real face samples from CelebA [64] as StyleGAN [29] was initially trained on a subset of it (the CelebA-HQ [28]) but also it was used as the seed dataset to re-train StyleGAN in section 3.3. While measuring the performance of the CelebA [64] dataset using a SoA FR tool, some irregularities were noticed. Investigating the matter, we find out that the CelebA dataset [64] has mislabelled samples (in terms of their identity), showing once more the drawbacks of real datasets and how difficult is to collect and correctly label them. This inspired the work described in section 3.4, in which a semi-automatic technique is proposed to find the mislabeled samples and correct the face dataset.

This work, apart from its general contributions of improving the quality of large real-world datasets, is essential in the process of creating the methodology for the validation of the synthetic samples, as without such clean 'real-world' datasets we cannot set correct baselines to make a fair comparisons with datasets built from synthetic data samples.

As we have prepared the tools to generate face samples (section 3.3) and have a clean distribution of the real samples that will be used to set a performance baseline (section 3.4), we are able to develop a methodology that will allow us to determine whether the generated face samples behave as the real face samples. More specifically, the goal is to determine if individual synthetic samples are unique in terms of identity, firstly with respect to the seed dataset that trains the StyleGAN model [29] and secondly with respect to other synthetic

face samples. In the research work presented in section 3.5, two approaches are introduced to enable the comparative analysis of large sets of synthetic face samples, for both cases. Furthermore, new metrics are introduced, and a technique is provided to remove the most connected data samples within a large synthetic dataset. Thus, the remaining synthetic samples can be considered as unique as data samples gathered from different real individuals. Through our experiments it is concluded that the resulting synthetic data samples exhibit excellent uniqueness when compared with the original training dataset, thus the generated samples can be used freely without any privacy issues, as it was validated that no generated face sample share the same identity attribute with a face sample from the seed dataset. Although when comparing the synthetic samples with one another, the proposed approaches showed that synthetic samples exist that share the same identity. Nevertheless, using the proposed methodology it is possible to remove the highly connected synthetic data samples. Thus, remaining with unique synthetic face samples in terms of their identity. The proposed methodology presented in section 3.5, is implemented using StyleGAN models [29, 36], although it can be implemented with any GAN model, trained on the task of face generation.

This work validates the synthetic data in the context of facial biometrics, as the proposed methodology determines whether or not the synthetic face samples behave like the real ones in terms of their identity attribute. Implementing this methodology, unique (in terms of identity) samples can be selected from a set of generated synthetic face samples, that can serve as the seed dataset. These seed samples can be further augmented, introducing variations such as light, pose, expression in order to complete the ultimate goal of building a large synthetic face dataset.

Furthermore, this work clearly indicates that synthetic samples should not be used arbitrarily, showing the importance of creating methods/metrics to validate the synthetic samples before incorporating them into ML/DL pipelines.

Overall, the works of this Chapter (3) present a framework on how to develop methods that validate synthetic data in the context of facial biometrics. The reasoning for selecting the category of facial biometrics, (which is also discussed in 3.1), is the large number of available datasets, the advanced AI tools related to that topic but also because facial data are susceptible to GDPR and privacy regulations. Now, this framework despite being focused on the category of facial biometrics, the same framework can be used to develop validation methodologies for synthetic samples for any data category and/or use case. Analysing the sections 3.3-3.5 the main steps can be summarised as following:

- The starting point is to build and understand the tools selected that generate synthetic data.

- Select the characteristic(s), based on which the "behaviour" of synthetic data will be examined.

- Select/Create the tools and metrics that will be used to measure the "behaviour" of the synthetic data in terms of the selected characteristic(s).

- Establish a baseline (based on real-world samples), which will be used to determine whether or not the synthetic samples follow the desired "behaviour". As explained earlier and shown from sections 3.4, it is important to have clean and correct real data samples, so that the baseline represents correctly the distribution of the real world.

- Finally, use/develop methods to identify and remove synthetic samples that do not behave as desired. In that way, it is ensured that remaining samples are validated and incorporate the desired characteristics so that they can be further used in the data pipelines.

Chapter 4

Utilising Synthetic Data in Training Pipelines for Edge-AI Solution

In this section an overview of the research work: *Varkarakis, Viktor, Wang Yao, and Peter Corcoran. "Towards End-to-End Neural Face Authentication in the Wild–Quantifying and Compensating for Directional Lighting Effects." arXiv preprint arXiv:2104.03854 (2021).*, [38], is given along with its research objectives and a discussion of contributions. The copy of the paper published based on this section is presented in the Appendix G. The contributions of its author related to the four major criteria as explained in section 1.4, for this research work [38] are presented in Table 4.1.

Table 4.1 Authors' Contributions to [38]

Contribution Criteria	*Contribution Percent*
Research Hypothesis/ Idea	VV 90%, PC 10%
Experiments and Implementation	VV 80%, WY 20%
Background	VV 80%, WY 20%
Manuscript Preparation	VV 80%, WY 10%, PC 10%

In the recent years face recognition has been well-studied in the literature with the most recent enhancements being driven by advances in CNN and deep learning [66, 68, 80, 81, 79]. Despite the improvements , factors such as pose [82–85], illumination [86–88], facial expression [89–91], age [92, 93] and, gender variations [94, 95], to name but a few, still affect the performance of the FR algorithms. In order to overcome these obstacles, in most of the literature the test samples for FR are assumed to be normalized in terms of their variations (pose, facial expression and illumination) to simplify the challenge of accurately distinguishing an

individual identity among a very large population.

The initial focus for implementation of neural algorithms, in embedded devices was on network optimizations such as parameter quantization and pruning, compressed convolutional filters and matrix factorizations [96]. However, the attention has recently shifted towards specialized neural topologies [97, 98] and ultra-low power realizations in hardware [99, 100]. These low power neural accelerators can perform ultra low power 'sensing', using SoA deep learning solutions, but these devices can't accommodate any processing in the central processing unit (CPU).

When considering the implementations of a SoA neural FR architectures in such low-power consumer appliances, the bottleneck is identified to be the pre-processing procedures required, that corrects the input facial sample before the image is fed to the FR network, due to the limitations of the low-power devices, that try to optimize the power consumption.

While it would be feasible to concatenate two (or more) ANN to allow a 'correction', followed by a neural FR solution, it requires a lot more parameters. Taking into account all possible variations (pose, illumination, expression, age, etc) that need correcting before a facial image is fed to the FR network, there could be a need for 3 or 4 'correcting' networks. Such solution will be extremely inefficient. As a result a better approach would be to fine-tune a single FR to generalise and be robust to all variations, eliminating the need for image prepossessing and correcting networks, optimising for low-power consumer appliances.

4.1   Research Objectives

In this work, our goal is to determine the feasibility of modifying a high accuracy SoA neural FR architecture to demonstrate robustness to un-normalised input image samples, in order to facilitate deployment in latest neural accelerators. This train of thoughts led to the research questions posed in this work [38]:

- Can we better quantify the effects of the external factors that affect fully neural FR and develop metrics to evaluate these;

- Can a fully neural FR architecture be modified through tuning and/or re-training to compensate internally for such external factors?

The research questions are generalised to include a range of external factors (e.g.: pose, illumination, expression, accessories, etc.) but in this work we focus on the effect of different

illumination scenarios. Specifically for the illumination variation, numerous image pre-processing methods exist to improve the performance of the FR model [101, 87, 102] but studies exploring the tuning or training of FR models to be robust to illumination variations, which are better suited for the modern neural accelerators are relatively rare [103, 104].

4.2   Summary of Contributions

In order to answer the research questions of this work with respect to the illumination variation, a SoA re-lighting technique is employed, known as Deep Single Image Portrait Relighting (DPR) [105] to augment a set of high-quality facial images with illumination from different directions (Appendix G, section 3.1). The effect of the of each illumination variation, on the performance of a SoA FR method is quantified using ROC curve techniques, and showed that a fully end-to-end neural FR solutions will be challenged by in-the-wild lighting conditions (Appendix G, section 4). Note that a re-lighting augmentation approach was adopted as existing public datasets do not provide sufficient lighting variability, which is discussed in the paper presented in the Appendix G, section 4.3.

Furthermore, the feasibility of handling lighting variations by fine-tuning the neural FR network is explored. The experiments demonstrate that using the augmented samples for fine-tuning, the FR model is able to recover to performance levels close to the original baseline for such illumination conditions. The fine-turning process also indicated that generalization from the primary directions to combinations of directional illumination is achieved - a promising result given the non-linear nature of lighting condition (Appendix G, section 5). Given these results it is clear that a full end-to-end neural FR optimised for implementation in the latest neural accelerators can be realized.

Instructions on how to generate the sets of CelebA-HQ [28] with the different illumination scenarios, a list of the images used in the experiments and the final fine-tuned FR network can be found in the GitHub repository of this work [6].

4.3   Discussion of Contributions

In this work [38], the lack of real world data for a known problem can be observed, as there aren't many real face samples with enough illumination variation in the literature (Appendix G, section 4.3). Capturing facial data in-the-wild, with precise illumination direction is a challenging task. Through this work [38], the effective use of advanced AI tools is shown, in

---

[6]Code available at:https://github.com/C3Imaging/Deep-Learning-Techniques/tree/ Quantify-Retrain-FR-for-Light

order to overcome the unavailability of real-world samples and enable the investigation of the task in hand. This work, utilises a SoA re-lighting technique to augment face samples with different illumination scenarios, without introducing any artifacts to the images (Appendix G, section 3.1). Using such AI tools, allows to have control over the image and be able to augment the samples accurately. This allows the study of the effect of each illumination scenario on the performance of the FR algorithm.

Furthermore, as we have sufficient face samples with several illumination scenarios it is possible to explore the possibility of fine-tuning the FR model, in order to be robust in such conditions. Our experiments showed that the fine-tuned FR model can cope with different illumination variations when trained with sufficient data that represents the problem correctly (Appendix G, section 5). Thus eliminating the need for pre-processing methods that normalise the lighting, and optimising such FR models, for easier deployment in latest neural accelerators.

Finally, the same framework presented in this work can be used to analyze the effect of other known factors (such as pose, expression, etc.) or/and their combination on the performance of the FR model but also train the FR model to be robust to these variations, so that pre-processing techniques can be eliminated from the FR pipeline.

Chapter 5

Contributions to Training Methodologies for Conditional GANs

In this section an overview of the research work: *Bazrafkan, Shabab, Viktor Varkarakis, Joseph Lemley, Hossein Javidnia, and Peter Corcoran. "Versatile Auxiliary Classification and Regression With Generative Adversarial Networks." IEEE Access 9 (2021): 38810-38825.*,[39] is given along with its research objectives and a discussion of contributions. This work brings together three unpublished works [106–108] that were initially presented as independent ,and organises them as a unified research piece. The copy of the paper published based on this section is presented in the Appendix H. The contributions of its author related to the four major criteria as explained in section 1.4, for this research work [39] are presented in Table 5.1. An addition section is given, where my personal motivation and contributions to this work are specified.

Table 5.1 Authors' Contributions to [39]

Contribution Criteria	Contribution Percent
Research Hypothesis Idea	SB 100%
Experiments and Implementation	SB 60%, VV 30%, HJ 10%
Background	SB 70%, JL 15%, VV 15%
Manuscript Preparation	SB 60%, VV 30%, PC 10%

There are several extensions to the original GAN idea [56]. One of them, known as Conditional GAN, where the original GAN is adapted so that the generator produces samples constrained to a specific class/aspect. Training them is one of the most appealing applications of GANs. With the recent restrictions on data acquisition and manipulation such as GDPR, it is become more important to be able to expand the existing databases and/or generate a new

set of anonymous samples from scratch. The conditional generators give the opportunity to construct data, given a specific class label. Some of the most successful implementations of conditional GANs include Conditional GAN (CGAN) [109], Auxiliary Classifier GAN (ACGAN) [110]. CGAN [109], achieves the conditioning of the generator's output by partitioning the latent space and also the auxiliary knowledge of the data class while in ACGAN [110] the loss of the CGAN is manipulated by adding a classification term which back-propagates through generator and discriminator.

5.1   Research Objectives

Despite the success of the Conditional GANs, some of these models come with various disadvantages as some can be versatile to be used with any GAN structure but there is no mathematical proof showing that the trained generator is able to provide distinct samples for different classes [110, 111]. Additionally, the classification / regression term is restrained to the discriminator's structure [110, 111], and therefore cannot be used with any GAN structure. Thus, limiting the potential for improvements, as a more optimal classification/regression network cannot be selected based on different tasks and relying on one structure for multiple problems. In this work, we seek new methods for training Conditional GANs, which will improve some of the disadvantages discussed.

5.2   Summary and Discussion of Contributions

The contribution of this work includes two new training methods of Conditional GANs for classification and regression tasks, which improve the disadvantages discussed earlier. The first method is called Versatile Auxiliary Classifier with Generative Adversarial Network (VAC+GAN) applied for classification tasks (binary and multi-class). The initial idea of VAC+GAN is then extended to regression tasks. This approach is called Versatile Auxiliary Regression with Generative Adversarial Network (VAR+GAN). The main idea of these methods is to remove the classification/ regression term from the discriminator's loss function, by adding a classification /regression network that back-propagates through the generator. Also, in the VAR+GAN scheme a new loss function is also proposed. Furthermore, the mathematical proofs for both methods are provided to show the validity and applicability of the proposed methods regardless of the GAN's and classifier's /regressor's structure or/and loss function (Appendix H,section III). The proposed schemes are applied in the experiments section for binary (classification), multi-class (classification) and regression problems, showing that using the proposed methods of VAC+GAN and VAR+GAN resulted

in superior results when compared with similar SoA techniques (Appendix H, section IV). The implementation of these methods can be found in the GitHub repository of this work at [7].

The main benefit of these methods is their versatility. The proposed methods [39] can be applied to any GAN structure with any loss function, as well as having the advantage of choosing any architecture for the classification/ regression network, opposed to other popular Conditional GAN schemes, where there is no flexibility in selecting the aforementioned components, as the classification / regression term is restrained to the discriminator's structure methods. This give us the ability, to choose the optimal components and optimise our approach depending on the investigated problem, rather than depend on a predefined scheme for a variety of problems. In our work [39], the proposed approaches are also validated through mathematical proofs and not just through experimental results.

5.3   Motivation and Personal Contributions

As discussed in 1.1, there is a need to build and/or improve the tools that generate synthetic data and/or augment these samples. This could be achieved either using advanced augmentation techniques, GANs and/or computer simulations. More specifically, the use of conditional GANs was investigated, for our use case of generating a large synthetic face dataset as discussed in Chapter 3. In our group (Cognitive, Connected & Computational Imaging Research - C3I) significant work has been done regarding augmentation techniques [112] but also GAN related works, in particular, the works [106–108], which focus on improving Conditional GANs. These [106–108] are part of a cohesive work that remained incomplete. In addition the initial implementation was in Theano on top of Lasagne, which is currently outdated and cannot be used easily. Being interested in incorporating these methods as part of my pipeline in creating a large synthetic facial dataset, I contributed into completing this work. More specifically my contribution to this research work are:

- Re-writing different parts and merging the three initial works [106–108] into a unified journal paper

- Being the corresponding author for the submission process

- Re-implementing the proposed methods in a current framework (PyTorch). The updated implementation of this work can be found at [7].

---

[7]Code available at: https://github.com/C3Imaging/Deep-Learning-Techniques/tree/vac_var_gan_pytorch

Through this work I had the opportunity to collaborate and be mentored from Shabab Bazrafkan, first author of the work from [39]. This experience was useful in being able to collaborate/ mentor the following Ph.D. students in continuing the research work of the CI3 group (which is discussed in 7.3).

Chapter 6

Additional Contributions

In this Chapter, some of my secondary publications are briefly mentioned. *Deep Learning for Consumer Devices and Services* is a series of articles in which the main focus of the discussion is on the deployment of deep learning on consumer devices. This series has currently four articles, from which I was involved in the following works:

- *Corcoran, Peter, Joseph Lemley, Claudia Costache, and Viktor Varkarakis. "Deep Learning for Consumer Devices and Services 2—AI Gets Embedded at the Edge." IEEE Consumer Electronics Magazine 8, no. 5 (2019): 10-19*, [40], which can be found in the Appendix I.

- *Corcoran, Peter, Claudia Costacke, Viktor Varkarakis, and Joseph Lemley. "Deep learning for consumer devices and services 3—Getting more from your datasets with data augmentation." IEEE Consumer Electronics Magazine 9, no. 3 (2020): 48-54*, [41], which can be found in the Appendix J.

In [40], a discussion is presented regarding the shift of AI from the data center into CE devices—"Edge-AI", practical use cases and as well the challenges of getting these deep learning solutions into CE devices. In [41], some of the challenges that data acquisitions entailed are presented. Furthermore an initial introduction of basic and advanced augmentation techniques is given, along with their benefits. In both works [40, 41], practical examples are borrowed from the research works of [33, 32] which are described in Chapter 2. These works relate to themes of the articles, as in [33, 32], advanced augmentation techniques are proposed to overcome the unavailability of data and contribute to a lower-complexity deep learning solution targeted for deployment in edge-devices .

Despite the fact that, these papers did not propose a new algorithm or methodology, they have been written in an accessible way and helped to introduce new researchers to the

subjects discussed. Such articles help promote and explain deep learning related themes to a broader audience.

Chapter 7

Conclusions and Future Works

In this section a concise summary of the outcomes for the main contributions of this thesis is given along with future works. Finally, a subsection with a 'reflective' discussion is presented with my personal thoughts related to the contributions of this thesis and the next steps.

7.1 Contribution to Advanced Augmentation Techniques

In Chapter 2, advanced augmentation techniques are proposed that transform high-quality frontal iris samples to a corresponding set of off-axis data samples. Using the proposed augmentations, it is possible to overcome the unavailability of data for the task of iris segmentation in AR/VR devices and generate a large number of images that represent the problem correctly. Thus allowing to train a CNN for this task effectively. As added benefits of using the proposed augmentation techniques, it is possible not only to increase the number of available samples but also inject variety and randomness to the new data samples. Utilising these components allowed us to train a low complexity network while keeping the performance comparable to the higher complexity, SoA SPDNN [25]. Thus, showing the importance of appropriate data augmentation strategies over the selection of heavy-weight network structures, that allows to design solutions that can be efficiently installed in edge devices, without losing significant performance (Appendix A).

Future work, with respect to the work presented in Chapter 2, will focus on refinements in the network design and training/augmentation methodologies to improve performance on specific AR/VR headsets. Some other practical examples of further research topics include developing an optimized CNN design based on SPDNN [25] methods with a similar, or perhaps even smaller number of parameters that can achieve similar segmentation accuracy

to the proposed network. Additional further work may include the use of the proposed augmentation techniques [32] for other applications where off-axis regions are noticed.

7.2   Contribution to Validation Methodologies for Synthetic Data

In Chapter 3, the idea of using AI tools to build large synthetic datasets and reduce/eliminate the need for real data samples is discussed. This is introduced in the work [34] which is presented in section 3.2. Swapping real data with synthetic is not straightforward, and therefore a framework/roadmap is proposed in [34](section 3.2), on the required steps in order to build and incorporate synthetic datasets effectively into our data/training pipelines. The first steps can be divided into two main categories:

- Building AI tools, that will generate a large amount of new data and/or augment these data samples.

- Create methodologies and techniques to validate that the generate data behave like real ones and also measure whether their use is effective when incorporated in the training pipelines.

Sections 3.3-3.5, contribute to the category of the validation of the synthetic data. A methodology is proposed that allows to determine whether or not the synthetic data behave as the real data, in the context of facial biometrics. More specifically, the identity attribute of synthetic face samples derived from GANs is explored. The proposed methodology allows to determine whether 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. Furthermore, techniques are proposed that are able to identify and remove the samples that aren't unique (in terms of their identity attribute), thus remaining with validated synthetic face samples that are considered as unique as data samples gathered from different real individuals.

The experiments conducted in [35], show that the generated data cannot be used arbitrarily in the data/training pipelines, highlighting the need for validation of synthetic data (Appendix F). Finally, despite the fact that the methodology proposed is applied in the context of the facial biometrics, a framework is given on how to develop methodologies that validate the synthetic data for any application (section 3.6).

In future works the methodology proposed in [35] can be used to generate a seed dataset of facial data samples that are validated and 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 [35], it would be interesting to investigate and benchmark different GANs for the task of generating face samples with a unique identity. Furthermore, utilizing core ideas from the proposed methodology [35], 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.

7.3   Utilising Synthetic Data in Training Pipelines for Edge-AI Solution

In Chapter 4, the lack of variation in terms of illumination scenarios in the current face datasets is observed and the effective use of a SoA relighting technique is explored to augment accurately face samples with different illumination directions. This allows to study the effect that the different illumination scenarios have on a SoA FR algorithm, with the experiments showing that the FR is greatly challenged in such conditions. Utilising the augmentation technique selected for this work, we are able to significantly increase the number of face samples with different illumination scenarios, which allows us to fine-tune the FR model. The experiments showed that when fine-tuned with the appropriate data, the FR model can be robust in different illumination conditions. Given that, the need for a pre-processing technique in the FR's pipeline, that usually normalises samples against illumination variations, is eliminated and therefore makes it easier for deployment in the latest neural accelerators (Appendix G). Thus showing how with the use of data augmentation it is feasible to develop end-to-end solutions that are preferred for installment in neural accelerators or/and edge devices.

Future works include using the same framework to conduct a broader study on factors that can affect the FR and investigate the potential of an end-to-end solution which can be robust to the different factors (such as pose, illumination, expression, etc.), without the need for pre-processing methods. As a continuation of this work [38], I am collaborating/ mentoring a Ph.D. student, who is investigating the effect of pose variation on the FR's performance, but also the pose and illumination combined, using the framework proposed in [38].

7.4   Contributions to Training Methodologies for Conditional GANs

In Chapter 5, the proposed training methods improve upon the drawbacks of conditional GANs for classification and regression cases (Appendix H). Conditional GANs are an AI tool that can be used in the GADAFAI framework [34] (Chapter 3), as they generate new samples, with the added benefit that they are also labelled. The main benefit of these methods is their versatility. The proposed methods can be applied to any GAN structure with any loss function, as well as having the advantage of choosing any architecture for the classification/ regression network, opposed to other popular Conditional GAN schemes, where there is no flexibility in selecting the aforementioned components, as the classification / regression term is restrained to the discriminator's structure methods. This give us the ability, to choose the optimal components and optimise our approach depending on the investigated problem, rather than depend on a predefined scheme for a variety of problems.

Future works includes extending the implementation for images of higher resolution and experiment with more GAN architectures. Additional, future studies involve, mixing the VAC+GAN and VAR+GAN ideas [39] to train conditional generators which accept discrete and continuous aspects at the same time. For example, in the face generation case one can generate random faces which belong to a specific gender class and also are fixed onto a particular landmark set.

7.5   Reflection of the Contributions and Next Steps

During my Ph.D. I was fortunate to witness the growth in the research area of deep learning, and be part of it. This growth is attributed to the availability of massive data sets (Big Data), and the arrival of new GPU-based hardware that enables these large data sets to be processed in reasonable timescales. Initially the approach adopted by researchers to deep learning can be characterised as model-centric, in which a fixed set of data is provided, and researchers iterate on the code/model to improve the performance of the deep learning algorithm [113]. As a product of this approach, SoA neural network architectures were developed such as ResNet, VGG, Inception and others [114–116]. As deep learning surpassed human performance [6, 7], it became more and more popular and was applied in a variety of fields yielding outstanding results in different machine learning applications [8–18].

While being a Ph.D. candidate, I worked closely with the industry partner of my program, Xperi, in solving industry problems in consumer technologies with the use of deep learning. Through my collaboration with Xperi and the research that I conducted, some practical realisation of deep neural network based solutions became more and more apparent. One of the biggest lesson was to realize the importance of data in the deep learning pipelines. I realised that the data should not be a considered an auxiliary part (of the deep learning pipelines) which just helps the network to learn but in fact that the network should serve the data. As a result, I understood that a neural network solution is only as good as the data used to train it and that the training data can determine how big the network should be, what kind of layers should be used, what non-linearity will serve better and what loss function should be employed in the model.

As a result of the realisations of the importance of data in deep learning pipelines, the drawbacks of real-world data and the difficulties associated with data acquisition (discussed in the Introduction, Chapter 1.1), this thesis is contributing to both data augmentation techniques and synthetic data as they can alleviate some of the drawbacks of real-world samples, eliminate/reduce the need for real-world samples and avoid the challenges related to the data acquisition process. Data augmentation techniques are proposed that transform the available data in order to simulate the image characteristics, for a custom problem, based on the a new camera setup. Utilising the augmentation techniques, enough samples are generated that represent the problem correctly, eliminating the need for a data acquisition process and allow to train effectively a CNN solution. With regards to the synthetic data, this thesis contributes by building AI tools, that will generate a large amount of new data and/or augment these data samples, so that we can eliminate the need for the real-world samples. Last but not least, this thesis contributes to creating methodologies validating that the generate data behave like real ones and also measure whether their use is effective when incorporated in the deep learning training pipelines.

Currently experts call for a broad shift to a more data-centric approach to deep learning, in which the code is fixed, and researchers are asked how to change or improve the data to improve performance [113] and the contributions of this dissertation aligns with this shift to a data-centric approach to deep learning. As a result of this shift , I expect AI tools that generate samples to be improved closing the gap to the real data and their use in the deep learning training pipelines to be increased.

Finally, I expept the development of deep learning solutions for sensors that can utilise the synthetic data to their full capacity, despite some of their drawbacks – such as realism (synthetic data can appear unrealistic, especially for data generated from current computer simulations). A good example is deep learning solutions for neuromorphic cameras, as these cameras do not capture images using a shutter as conventional cameras do [117]. Instead, each pixel inside an event camera operates independently and asynchronously, reporting changes in brightness as they occur, and staying silent otherwise [117]. Thus in such a case synthetic data can be very useful as the most important characteristic of the samples captured from a neuromorphic sensor is the changes that occur in each pixel rather than the whole image.

References

[1] M. Minsky and S. Papert, "An introduction to computational geometry," *Cambridge tiass., HIT*, vol. 479, p. 480, 1969.

[2] "Deep learning," https://en.wikipedia.org/wiki/Deep_learning, accessed: 2022-01-22.

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[4] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[6] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of big data*, vol. 2, no. 1, pp. 1–21, 2015.

[9] G. Dahl, M. Ranzato, A.-r. Mohamed, and G. E. Hinton, "Phone recognition with the mean-covariance restricted boltzmann machine," *Advances in neural information processing systems*, vol. 23, pp. 469–477, 2010.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[11] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[12] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 14–22, 2011.

[13] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[14] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[16] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocký, "Empirical evaluation and combination of advanced language modeling techniques," *Proceedings of Interspeech*, pp. 605–608, 01 2011.

[17] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection." in *NIPS*, vol. 24, 2011, pp. 801–809.

[18] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 127–135.

[19] C. H. Martin, T. S. Peng, and M. W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.

[20] J. Shijie, W. Ping, J. Peiyi, and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," in *2017 Chinese automation congress (CAC)*. IEEE, 2017, pp. 4165–4170.

[21] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[22] F. J. Moreno-Barea, J. M. Jerez, and L. Franco, "Improving classification accuracy using data augmentation on small data sets," *Expert Systems with Applications*, vol. 161, p. 113696, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305200

[23] "Automating data augmentation: Practice, theory and new direction," http://ai.stanford.edu/blog/data-augmentation/, accessed: 2022-01-23.

[24] S. Bazrafkan, "Contributions to deep learning methodologies," Ph.D. dissertation, NUI Galway, 2018.

[25] S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end deep neural network for iris segmentation in unconstrained scenarios," *Neural Networks*, vol. 106, pp. 79–95, 2018.

[26] A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira, "Mobbio: A multimodal database captured with a portable handheld device," in *2014 International conference on computer vision theory and applications (VISAPP)*, vol. 3. IEEE, 2014, pp. 133–139.

[27] "Synthetic data," https://en.wikipedia.org/wiki/Synthetic_data, accessed: 2022-01-23.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[30] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv preprint arXiv:1909.11512*, 2019.

[31] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[32] V. Varkarakis, S. Bazrafkan, and P. Corcoran, "Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets," *Neural Networks*, vol. 121, pp. 101–121, 2020.

[33] ——, "A deep learning approach to segmentation of distorted iris regions in head-mounted displays," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 2018, pp. 1–9.

[34] P. Corcoran, H. Javidnia, J. E. Lemley, and V. Varkarakis, "Generative augmented dataset and annotation frameworks for artificial intelligence (gadafai)," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.

[35] V. Varkarakis, S. Bazrafkan, G. Costache, and P. Corcoran, "Validating seed data samples for synthetic identities–methodology and uniqueness metrics," *IEEE Access*, vol. 8, pp. 152 532–152 550, 2020.

[36] V. Varkarakis, S. Bazrafkan, and P. Corcoran, "Re-training stylegan-a first step towards building large, scalable synthetic facial datasets," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.

[37] V. Varkarakis and P. Corcoran, "Dataset cleaning—a cross validation methodology for large facial datasets using face recognition," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[38] V. Varkarakis, W. Yao, and P. Corcoran, "Towards end-to-end neural face authentication in the wild–quantifying and compensating for directional lighting effects," *arXiv preprint arXiv:2104.03854*, 2021.

[39] S. Bazrafkan, V. Varkarakis, J. Lemley, H. Javidnia, and P. Corcoran, "Versatile auxiliary classification and regression with generative adversarial networks," *IEEE Access*, vol. 9, pp. 38 810–38 825, 2021.

[40] P. Corcoran, J. Lemley, C. Costache, and V. Varkarakis, "Deep learning for consumer devices and services 2—ai gets embedded at the edge," *IEEE Consumer Electronics Magazine*, vol. 8, no. 5, pp. 10–19, 2019.

[41] P. Corcoran, C. Costacke, V. Varkarakis, and J. Lemley, "Deep learning for consumer devices and services 3—getting more from your datasets with data augmentation," *IEEE Consumer Electronics Magazine*, vol. 9, no. 3, pp. 48–54, 2020.

[42] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, "Beyond authorship: attribution, contribution, collaboration, and credit," *Learned Publishing*, vol. 28, no. 2, pp. 151–155, 2015.

[43] J. Lemley, "Deep learning techniques in data augmentation and neural network design," Ph.D. dissertation, NUI Galway, 2020.

[44] B. Kress, E. Saeedi, and V. Brac-de-la Perriere, "The segmentation of the hmd market: optics for smart glasses, smart eyewear, ar and vr headsets," in *Photonics Applications for Aviation, Aerospace, Commercial, and Harsh Environments V*, vol. 9202. International Society for Optics and Photonics, 2014, p. 92020D.

[45] M. Linao, "The present and future of vr/ar: Applications in education, gaming, commerce, and industry," *CB Insights*, 2016.

[46] T. Economist, "The promise of augmented reality," 2018.

[47] K. W. Ching and M. M. Singh, "Wearable technology devices security and privacy vulnerability analysis," *International Journal of Network Security & Its Applications*, vol. 8, no. 3, pp. 19–30, 2016.

[48] D. K. Yadav, B. Ionascu, S. V. K. Ongole, A. Roy, and N. Memon, "Design and analysis of shoulder surfing resistant pin based authentication mechanisms on google glass," in *International conference on financial cryptography and data security*. Springer, 2015, pp. 281–297.

[49] S. Thavalengal, P. Bigioi, and P. Corcoran, "Efficient segmentation for multi-frame iris acquisition on smartphones," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2016, pp. 198–199.

[50] H. Proença and L. A. Alexandre, "Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage," *Image and vision computing*, vol. 28, no. 1, pp. 202–206, 2010.

[51] H. Hofbauer, F. Alonso-Fernandez, J. Bigun, and A. Uhl, "Experimental analysis regarding the influence of iris segmentation on the recognition rate," *Iet Biometrics*, vol. 5, no. 3, pp. 200–211, 2016.

[52] M. Erbilek, M. C. Da Costa-Abreu, and M. Fairhurst, "Optimal configuration strategies for iris recognition processing," 2012.

[53] "Casia iris image database," http://biometrics.idealtest.org/, accessed: 2021-11-26.

[54] S. Rakshit, "Novel methods for accurate human iris recognition," *University of Bath*, 2007.

[55] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, "The ubiris. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1529–1535, 2009.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[57] T. Silva, "An intuitive introduction to generative adversarial networks (gans)," https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/, January 2018.

[58] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–49, 2021.

[59] H. Chen, "Challenges and corresponding solutions of generative adversarial networks (gans): A survey study," in *Journal of Physics: Conference Series*, vol. 1827, no. 1. IOP Publishing, 2021, p. 012066.

[60] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[61] B. Shen, B. RichardWebster, A. O'Toole, K. Bowyer, and W. J. Scheirer, "A study of the human perception of synthetic faces," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[62] I. Goodfellow, "4.5 years of gan progress on face generation," https://twitter.com/goodfellow_ian/status/1084973596236144640?lang=en, January 2021.

[63] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[64] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[65] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[66] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[67] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[68] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[69] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[70] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th international conference on computer vision.* IEEE, 2009, pp. 365–372.

[71] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision.* Springer, 2016, pp. 87–102.

[72] I. Gallo, S. Nawaz, A. Calefati, and G. Piccoli, "A pipeline to improve face recognition datasets and applications," in *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ).* IEEE, 2018, pp. 1–6.

[73] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[74] S. Ding, J. Wu, W. Xu, and H. Chao, "Automatically building face datasets of new domains from weakly labeled data with pretrained models," *arXiv preprint arXiv:1611.08107*, 2016.

[75] C. Jin, R. Jin, K. Chen, and Y. Dou, "A community detection approach to cleaning extremely large face database," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[76] S. Bazrafkan, H. Javidnia, and P. Corcoran, "Latent space mapping for generation of object elements with corresponding data annotation," *Pattern Recognition Letters*, vol. 116, pp. 179–186, 2018.

[77] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs, "Deep single-image portrait relighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7194–7202.

[78] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.

[79] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[80] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[81] M. Mehdipour Ghazi and H. Kemal Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 34–41.

[82] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1415–1424.

[83] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2207–2216.

[84] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and machine intelligence*, vol. 19, no. 7, pp. 743–756, 1997.

[85] S.-I. Choi, C.-H. Choi, and N. Kwak, "Face recognition based on 2d images under illumination and pose variations," *Pattern Recognition Letters*, vol. 32, no. 4, pp. 561–571, 2011.

[86] J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, and P. J. Phillips, "Quantifying how lighting and focus affect face recognition performance," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*.   IEEE, 2010, pp. 74–81.

[87] J.-Y. Zhu, W.-S. Zheng, F. Lu, and J.-H. Lai, "Illumination invariant single face image recognition under heterogeneous lighting condition," *Pattern Recognition*, vol. 66, pp. 313–327, 2017.

[88] J.-W. Wang, N. T. Le, J.-S. Lee, and C.-C. Wang, "Illumination compensation for face recognition using adaptive singular value decomposition in the wavelet domain," *Information Sciences*, vol. 435, pp. 69–93, 2018.

[89] M. Pavlovic, R. Petrovic, B. Stojanovic, and S. Stankovic, "Facial expression and lighting conditions influence on face recognition performance," in *Proceedings of the 5th International Conference IcETRAN*, 2018, pp. 777–781.

[90] A. Peña, A. Morales, I. Serna, J. Fierrez, and A. Lapedriza, "Facial expressions as a vulnerability in face recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*.   IEEE, 2021, pp. 2988–2992.

[91] Y. Peng and H. Yin, "Facial expression analysis and expression-invariant face recognition by manifold-based synthesis," *Machine Vision and Applications*, vol. 29, no. 2, pp. 263–284, 2018.

[92] S. Riaz, Z. Ali, U. Park, J. Choi, I. Masi, and P. Natarajan, "Age-invariant face recognition using gender specific 3d aging modeling," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 25 163–25 183, 2019.

[93] D. Deb, L. Best-Rowden, and A. K. Jain, "Face recognition performance under aging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 46–54.

[94] N. Narang and T. Bourlai, "Gender and ethnicity classification using deep learning in heterogeneous face recognition," in *2016 International Conference on Biometrics (ICB)*.   IEEE, 2016, pp. 1–8.

[95] C. Werther, M. Ferguson, K. Park, T. Kling, C. Chen, and Y. Wang, "Gender effect on face recognition for a large longitudinal database," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[96] A. Goel, C. Tung, Y.-H. Lu, and G. K. Thiruvathukal, "A survey of methods for low-power deep learning and computer vision," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*. IEEE, 2020, pp. 1–6.

[97] Z. He, B. Gong, and D. Fan, "Optimize deep convolutional neural network with ternarized weights and high accuracy," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 913–921.

[98] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *arXiv preprint arXiv:1605.04711*, 2016.

[99] B. Fleischer, S. Shukla, M. Ziegler, J. Silberman, J. Oh, V. Srinivasan, J. Choi, S. Mueller, A. Agrawal, T. Babinsky *et al.*, "A scalable multi-teraops deep learning processor core for ai trainina and inference," in *2018 IEEE Symposium on VLSI Circuits*. IEEE, 2018, pp. 35–36.

[100] W. Guicquero and A. Verdant, "Algorithmic enablers for compact neural network topology hardware design: Review and trends," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[101] J. Ruiz-del Solar and J. Quinteros, "Illumination compensation and normalization in eigenspace-based face recognition: A comparative study of different pre-processing approaches," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1966–1979, 2008.

[102] A. R. Shinwari, A. J. Balooch, A. A. Alariki, and S. A. Abdulhak, "A comparative study of face recognition algorithms under facial expression and illumination," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2019, pp. 390–394.

[103] D. Crispell, O. Biris, N. Crosswhite, J. Byrne, and J. L. Mundy, "Dataset augmentation for pose and lighting invariant face recognition," *arXiv preprint arXiv:1704.04326*, 2017.

[104] H. A. Le and I. A. Kakadiaris, "Illumination-invariant face recognition with deep relit face images," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 2146–2155.

[105] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs, "Deep single-image portrait relighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[106] S. Bazrafkan, H. Javidnia, and P. Corcoran, "Versatile auxiliary classifier with generative adversarial network (vac+ gan)," *arXiv preprint:1805.00316*, 2018.

[107] S. Bazrafkan and P. Corcoran, "Versatile auxiliary classifier with generative adversarial network (vac+gan), multi class scenarios," *arXiv preprint:1806.07751*, 2018.

[108] ——, "Versatile auxiliary regressor with generative adversarial network (var+ gan)," *arXiv e-prints*, pp. arXiv–1805, 2018.

[109] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[110] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[111] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 440–448.

[112] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[113] A. NG, "Mlops: From model-centric to data-centric ai," https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf, 2022, accessed: 2022–01-24.

[114] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[115] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[116] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[117] "Event camera," https://en.wikipedia.org/wiki/Event_camera, accessed: 2022-01-24.

Appendix A

Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets
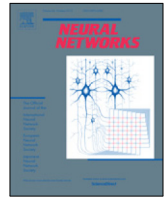
Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets

Viktor Varkarakis [a,*], Shabab Bazrafkan [b], Peter Corcoran [a]

[a] *Department of Electronic Engineering, College of Engineering, National University of Ireland Galway, University Road, Galway, Ireland*
[b] *imec-Vision Lab, Department of Physics, University of Antwerp, Antwerp, Belgium*

ABSTRACT

A data augmentation methodology is presented and applied to generate a large dataset of off-axis iris regions and train a low-complexity deep neural network. Although of low complexity the resulting network achieves a high level of accuracy in iris region segmentation for challenging off-axis eye-patches. Interestingly, this network is also shown to achieve high levels of performance for regular, frontal, segmentation of iris regions, comparing favourably with state-of-the-art techniques of significantly higher complexity. Due to its lower complexity this network is well suited for deployment in embedded applications such as augmented and mixed reality headsets.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Data augmentation is a common technique in Deep Learning and is frequently exploited by researchers in order to overcome the obstacle of limited labelled data. But even where large datasets are available using appropriate data augmentation techniques can improve the distribution of training samples and reduce overfitting during training. In turn this increases the generalization of the trained network and improves network accuracy and robustness.

Commonly used data augmentation techniques involve rotation, translation, flipping, re-sizing or affine transformation of data samples. Other well known techniques include adding noise, motion or optical blur, or varying image contrast or gamma. These techniques are computationally inexpensive (Krizhevsky, Sutskever, & Hinton, 2012) and have been previously used successfully to reduce overfitting in training a CNN for the *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) (Russakovsky et al., 2015), and achieved state-of the-art results at that time.

Data augmentation methods have been widely used in deep learning, and the selection of appropriate data augmentation strategies can be more important to solving a machine learning problem than the choice of a particular neural-network structure (Goodfellow, Bengio, & Courville, 2016; Shijie, Ping, Peiyi, & Siping, 2017).

In this work we focus on presenting and validating a new data-augmentation technique targeted to optimize the performance of off-axis iris segmentation. Our results show promising levels of accuracy for off-axis segmentation and the resulting trained neural network has performance that is competitive with state-of-art frontal iris segmentation networks of much greater complexity. This illustrates the powerful generalizing capabilities of our augmentation methodology.

The core of this work takes high-resolution frontal iris datasets as a starting point for generating a corresponding set of off-axis data samples. A similar approach can be adopted and applied to create a set of off-axis data samples from any frontal object view, but in this work we focus on the problem of off-axis iris segmentation as this is a new, emerging challenge for user authentication on Augmented and Mixed reality headsets.

Biometric user authentication is available on consumer devices, including smartphones, using facial recognition (Darwaish, Moradian, Rahmani, & Knauer, 2014; Samangouei, Patel, & Chellappa, 2017; Vazquez-Fernandez & Gonzalez-Jimenez, 2016) and fingerprint biometric (Bakir, Chesler, & Torriente, 2016; Cherapau, Muslukhov, Arachchilage, & Beznosov, 2015; De Luca, Hang, von Zezschwitz, & Hussmann, 2015; Goode, 2014; Ring, 2015; Tipton, White II, Sershon, & Choi, 2014). The broad adoption of biometrics on consumer devices was originally discussed in Corcoran (2013) with additional discussion of the impacts in several following articles (Corcoran, 2016, 2017; Corcoran & Costache, 2016). Being a near ideal biometric, the iris of the human eye is well-suited to many consumer applications, but iris recognition is traditionally implemented in a controlled environment and under constrained acquisition conditions.

* Corresponding author.
E-mail addresses: v.varkarakis1@nuigalway.ie (V. Varkarakis), shabab.bazrafkan@uantwerpen.be (S. Bazrafkan), peter.corcoran@nuigalway.ie (P. Corcoran).

Authentication requirements in consumer devices are evolving beyond today's mobile devices. New virtual reality (VR) and augmented reality (AR) headsets provide a gateway to sophisticated virtual worlds and online services (Kress, Saeedi, & Brac-de-la Perriere, 2014; Linao, 2016; Timekeeper, 2017). In fact researchers have been working with Augmented Displays for more than 20 years (Bhorkar, 2017; Mann, 2001, 2004, 2013; Mann & Fung, 2002; Starner et al., 1997; Tang, Aimone, Fung, Marjan, & Mann, 2002). The most recent mass market experiment with a wearable, augmented/mediated-reality display, that could be worn on a day-to-day basis, was Google Glass (Ackerman, 2013; Hayes, 2016; Mann, 2013). *Glass*, as it became known, was considered to be a game changing technology for a few years across a wide range of industry sectors (Elise, 2014; Fox & Felkey, 2013; Muensterer, Lacher, Zoeller, Bronstein, & Kübler, 2014; Schreinemacher, Graafland, & Schijven, 2014). But ultimately, the product was withdrawn (Cave, 2015).

A key challenge with AR/VR headsets is that, lacking a physical keyboard they do not provide an intuitive mean of user authentication. The weak authentication available in *Glass* (Ching & Singh, 2016; Hayes, 2016; Yadav, Ionascu, Ongole, Roy, & Memon, 2015) subsequently led to various attempts to refine and improve on the basic authentication of the headset (Chan, Halevi, & Memon, 2015; Chauhan, Asghar, Kâafar, & Mahanti, 2016; Peng et al., 2017). Ultimately, the device authentication was simply not adequate and led, in part, to its withdrawal from the market.

This leads us to consider how the next generation of wearable AR/VR vision systems might implement a more seamless and intuitive authentication mechanism without sacrificing security and robustness. The implementation of a face recognition system is not practical, as the form-factor of an AR/VR head-set does not allow to capture a full facial image. However, with the reduction in size and cost of multi-cameras systems on mobile devices it is now practical to consider that rear-facing (i.e. user-facing) camera systems can be incorporated into such headsets. One important driver for rear-facing cameras is the use of eye-tracking to dynamically determine the wearer's point of gaze (PoG) which is important for accurate AR/VR rendering (Cognard, Goncharov, Devaney, Dainty, & Corcoran, 2018; Rompapas et al., 2017).

Iris authentication is a proven and reliable biometric trait with high distinctiveness, permanence and performance (Prabhakar, Pankanti, & Jain, 2003). The use of iris recognition on consumer devices is explored across multiple works (Corcoran, Bigioi, & Thavalengal, 2015; Shejin & Corcoran, 2016; Thavalengal, Andorko, Drimbarean, Bigioi, & Corcoran, 2015; Thavalengal, Bigioi, & Corcoran, 2015b) and the importance of accurate iris segmentation, particularly in consumer imaging devices, is identified as a key challenge (Bazrafkan, Thavalengal, & Corcoran, 2018; Thavalengal, Bigioi, & Corcoran, 2016). In the iris authentication workflow, failed segmentations represent the single largest source of error (Erbilek, Da Costa-Abreu, & Fairhurst, 2012; Hofbauer, Alonso-Fernandez, Bigun, & Uhl, 2016; Proença & Alexandre, 2010). In addition to its role in improving the performance of an iris-based authentication system, the accurate segmentation of iris regions, can be used successfully for eye-gaze estimation (Hammal, Massot, Bedoya, & Caplier, 2005). Eye-gaze as mentioned is a key element of various user-interface modalities for wearable AR/VR displays.

1.1. Background to the problem

Data augmentation techniques and their effectiveness in computer vision have been explored in multiple research works.

In Shijie et al. (2017) and Taylor and Nitschke (2017) the effect of several data augmentation techniques (GAN/WGAN, Flipping, Cropping, Shifting, PCA jittering, Colour jittering, Noise, Rotation)

is investigated, in improving the performance of a CNN in the image classification task. Furthermore, researchers have studied ways to implement and automate different data augmentation techniques (beyond the traditional techniques) that can boost a CNN's performance.

In Perez and Wang (2017) the authors present an approach called neural augmentation, which allows a neural network to learn augmentations that will improve the accuracy of a CNN classifier. Similarly in Cubuk, Zoph, Mane, Vasudevan, and Le (2018) where their Autoaugment method searches automatically for improved data augmentation policies in order for a neural network to yield higher performance in the validation set.

Additionally in Lemley, Bazrafkan, and Corcoran (2017), the Smart Augmentation method learns suitable augmentations during the training of the neural network. The results indicate that this method could be applied for a range of tasks and that by implementing this augmentation method a small network achieved better results than those obtained by a much larger network. Finally the effectiveness of data augmentation is not applicable only to computer vision problems but also in audio (Salamon & Bello, 2017; Schlüter & Grill, 2015). It is worthwhile to mention that these methods are solely designed for classification problems and data augmentation for regression problems are widely applied by prior knowledge.

As new technologies arise, deep neural networks can provide solutions to the consequently new challenges that appear, but the data collection related to the new problems and their annotation remains the bottleneck of the process. Data augmentation is commonly used to boost the performance and the generalization of the CNNs but with specialized techniques can be utilized to eliminate the expensive procedure of the data collection–annotation. By designing specific data augmentation techniques, it is possible to simulate the characteristics and features related to the new challenges and inject them in relation to the problem datasets, so that a CNN can be efficiently train for the new task without going through a data acquisition and annotation procedure.

In this case the new challenges arise from the next generation of wearable AR/VR vision systems which as mentioned in the introduction will have to implement a more seamless and intuitive authentication mechanism that is available on today's mobile devices. Therefore, the next generation of AR/VR headsets might incorporate a user-facing camera installed below the eye level, used for iris authentication or eye-gaze tracking. In such a case, off-axis iris images can be readily obtained and provide a suitable biometric for user authentication.

Note that a key challenge for accurate iris recognition is to accurately segment the iris region (Bazrafkan et al., 2018; Jillela & Ross, 2013). Given that camera locations for AR/VR devices must be mounted off-axis, often with an oblique perspective and close proximity to the observed eye-region the segmentation process for such off-axis iris regions becomes even more critical as errors at the segmentation stage are propagated to the feature extraction and pattern matching stages of the authentication workflow (Bazrafkan et al., 2018; Hofbauer et al., 2016; Hofbauer, Alonso-Fernandez, Wild, Bigun, & Uhl, 2014). While there are past studies on off-axis iris and its effects on recognition rates, the problem of near-eye iris segmentation is a new problem arising from the introduction of emerging AR/VR headset technology into consumer devices and a main obstacle for providing a solution, is the unavailability of a dataset with iris samples captured from an user-facing camera on AR/VR devices, but data augmentation techniques can be utilized to eliminate this obstacle.

Currently, the majority of existing iris recognition systems follow the authentication workflow as (i) image acquisition: an eye image is acquired using a camera, (ii) iris segmentation: eye/iris region is located in this image followed by isolating the region

representing iris. (iii) iris normalization (iv) Feature extraction: relevant features which represent the uniqueness of the iris pattern is extracted from the iris region and (v) similarity of the two iris representations is evaluated by pattern matching techniques. The described workflow is illustrated in Fig. 1, highlighting the focus of this work, which is on the second step of the iris recognition workflow, the iris segmentation. Specifically, specialized data augmentation techniques are designed to simulate the iris images as represented when captured from a user-facing camera on an AR/VR device along with their ground truth segmentation map and thus eliminating the expensive process of data acquisition and annotation. In continuance a low-complexity network is designed and trained with the augmented samples to successfully segment the off-axis close proximity iris images. Although this work is focused solely on the use of deep neural networks for the off-axis iris segmentation task, it could be widely applicable for other off-axis region segmentation problems.

1.2. Related literature

The significant success of deep neural networks at vision-oriented tasks has enabled exceptional advancement in semantic segmentation. An instance of that is the DeepLab method (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014), where instead of using deconvolution, they proposed Atrous ('Holes') convolution. The proposed method is combined with fully connected conditional random fields (CRF) and is able to produce semantically accurate predictions and detailed segmentation maps efficiently. In a follow-up publication (Chen, Papandreou, Kokkinos, Murphy and Yuille, 2018), the same team proposed an atrous spatial pyramid pooling (ASPP) module, consisting of multiple parallel atrous convolutional layers with different sampling rates to improve the segmentation of objects. The DeepLab team consistently proposes methods for improving the segmentation standards (Chen, Papandreou, Schroff, & Adam, 2017; Chen, Zhu, Papandreou, Schroff and Adam, 2018). In another semantic segmentation approach (Jiang, Yuan, & Wang, 2018), the authors develop an adaptive-depth neural network to obtain the course semantic segmentation results. At the same time, the contour information is provided by a contour-aware network. Both the coarse semantic information and contour information are modelled in the same way and combining this information the semantic labels are given through global inference based on CRF. Furthermore, semantic segmentation methods could provide a solution to road detection, a challenging problem in autonomous driving. In Wang, Gao, and Yuan (2018), a Siamese neural network is developed which accepts an RGB image, the semantic contour and the location prior, at the same time, for the road detection task and the results demonstrate that the network is able to learn discriminative features of road boundaries and location prior. More detailed information for semantic segmentation techniques using deep learning can be found in Garcia-Garcia, Orts-Escolano, Oprea, Villena-Martinez, and Garcia-Rodriguez (2017) and Lateef and Ruichek (2019).

As the work is focused on the iris segmentation task, a quick overview of similar works in the literature is outlined below.

1.2.1. Frontal iris segmentation

The number of methods in the literature regarding iris segmentation shows that this topic has been thoroughly studied but remains an active area of research. When referring to iris segmentation algorithms, a good starting point is two highly cited works in the literature: Daugman (2009) and Wildes (1997). In the iris matching algorithms developed in these research papers iris segmentation is achieved by fitting a circular contour to the iris and pupil. These two methods differ mostly in the way

they define the circular boundaries on the image information. Daugman's integrodifferential operator searches the entire image pixel by pixel to find the best circular path for the iris and pupil boundaries. While Wilde, in order to fit the circular contour, combines an edge detector and Hough transform.

In continuance approaches were implemented in an attempt to speed up the process. For example, Liu, Yuan, Zhu, and Cui (2003), uses a Canny edge detector with a Hough transform to provide a fast localization of the iris edges with the assumption that the iris texture is located between two homocentric circles. Several other methods were developed based on Wilde's and Daugman's implementations such as: Huang's (Huang, Luo, & Chen, 2002), Khan's (Khan et al., 2011), He's and Shi's (He & Shi, 2006), Lili's and Mei's (Lili & Mei, 2005).

As noted, the aforementioned methods assume the circularity of the iris outer boundary and pupil boundaries. However, Daugman in his follow up work (Daugman, 2007) shows that a non-circularity applies to the iris and pupil contour which when defined precisely has a significant influence on recognition performance. Therefore, adopts an active contour or snake model to segment the iris. Furthermore, Shah and Ross (2009) implemented a geodesic active contour to capture the iris texture and experimental result on non-ideal iris images designates the effectiveness of this method. Koh, Govindaraju, and Chaudhary (2010) similarly implemented an active contour model which was combined with the Hough transform for iris localization. In another approach, Broussard and Ives (2009), used a feature saliency algorithm to identify the measurements that could define the iris boundary. The selected measurements are fed to a shallow artificial neural network in order to accurately predict the outer iris boundary. A detailed overview of the iris segmentation literature can be found in Bowyer, Hollingsworth, and Flynn (2008, 2013), as well as in Jan (2017) where approaches for segmenting non-ideal iris images are reviewed.

1.2.2. Off-axis iris segmentation

A subsection of non-ideal iris images includes the off-axis iris images. Localizing the iris in this type of images has always been a challenge for researchers. In Dorairaj, Schmid, and Fahmy (2005), Dorairaj assumes that a rough estimation of the angle rotation is available in order to deal with the off-axis iris problem. Two different objective functions are used to refine the estimate. When two images are available from the same iris class, the "ideal" and off-axis iris image, the Hamming distance between the ICA coefficients of the two images is calculated. In the case that only the off-axis image is available, Daugman's integro-differential operator is used. A projective transformation is applied to rotate the off-axis image into a frontal view image once the angle is estimated. In the next step, the image is enhanced and segmented with the integro-differential operator. In another approach, Li in Li (2006) first fits an ellipse to the pupil boundaries. After that based on the information that has been retrieved from the ellipse fitting, rotation and scaling are applied to the image, to restore the straight position of the ellipse and the circularity of the pupil. The segmentation of iris is then operated by Daugman's like algorithms. A similar approach can be found in Abhyankar, Hornak, and Schuckers (2005) where the use of projective and affine transformation is explored in order to bring the off-axis iris images and match them with frontal iris images. This approach comes with some serious downsides, such as the blurring of the iris outer boundaries and the fact that a prior knowledge of the angle is required for the transformation. Finally, in Abhyankar and Schuckers (2006) the use of active shape models to retrieve the elliptical boundaries of the off-axis iris is investigated.
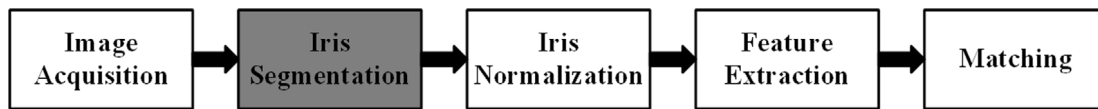
Fig. 1. Iris authentication workflow. In practical implementation the bulk of authentications errors are due to incorrect segmentations (Erbilek et al., 2012; Hofbauer et al., 2016; Proença & Alexandre, 2010).

1.2.3. Deep learning approaches for iris segmentation

Liu, in Liu et al. (2016) proposed two CNN approaches to segment noisy iris images acquired under unconstrained conditions. In the first approach called hierarchical convolutional neural networks (HCNNs), three patches taken from different scales of the same image are used as input. The HCNN consists of three similar blocks, a combination of convolutional and pooling layers that are merged together into a fully connected layer. In the second approach, 31 convolutional layers and 6 pooling layers are used to compose the multi-scale fully convolutional network (MFCNs). Both models are end-to-end, with no requirement for pre- or post-processing of the image. Arsalan et al. (2017), introduced a two-stage iris segmentation method. The first stage includes a pre-processing of the image and the use of a modified Hough Transform to identify the region of interest (ROI). In the second stage, a mask of [21 × 21] pixels, based on the ROI defined in the previous stage, is fed to a pre-trained VGG-face model which classifies the pixels as iris or non-iris. In a follow up work which is focused on segmenting low quality iris images, Arsalan in Arsalan et al. (2018), proposed a densely connected fully convolutional network (IrisDenseNet), consisting of two main components: a densely connected encoder and a SegNet decoder. In a similar work, Bazrafkan in Bazrafkan et al. (2018), presented a network design focused on segmenting iris of inferior quality. Four different end-to-end fully convolutional networks are merged into a single model using a method known as Semi Parallel Deep Neural Networks (SPDNN). In this way, the final model benefits from each of the four distinct network designs. Furthermore from a more medical aspect in Lakra, Tripathi, Keshari, Vatsa, and Singh (2018), utilizes a DenseNet-121 which has four convolution blocks to be able to segment iris with cataract or post cataract surgery. Finally, since the existence of a large labelled dataset is a prerequisite in order to implement a convolutional neural network approach, Jalilian in Jalilian, Uhl, and Kwitt (2017) to overcome this obstacle, introduced a domain adaption method so that a CNN for iris segmentation could be trained with a limited data.

1.3. Contributions

The focus of this work is to improve the segmentation of off-axis iris images originating from the unconstrained conditions of a user-facing camera on wearable AR/VR device.

The model proposed is an end to end deep neural network which accepts an off-axis eye-region image and generates the corresponding binary segmentation map for the iris region as output. Performance evaluation of the proposed model shows advantages over recent iris segmentation techniques in the literature which together with its simple, yet efficient design makes it well-suited for deployment in wearable AR/VR devices.

Three noteworthy contributions are presented in this work.

1. Specialized data augmentation methods that generate distorted iris images of size and quality typical of the user-facing camera employed on today's wearable AR/VR headsets. These are derived from a high-quality iris dataset together with a corresponding ground truth.

2. By utilizing the data augmentation techniques and producing a large number of representative data, it enabled us to propose an improved low complexity neural network design for the off-axis iris segmentation task with reduced memory and computational requirements in comparison with other deep learning state-of-the-art iris segmentation techniques while achieving equivalent performance.

3. A thorough evaluation of the proposed segmentation model is presented on several well-known public iris datasets. The presented method is compared with state-of-the-art iris segmentation techniques.

1.4. Foundation methods

1.4.1. Data augmentation

Despite the large amount of data available today, there are many problems where data collection and annotation poses challenges and thus only small datasets are publicly available. In some cases, the data contains sensitive information such as in medical applications or due to privacy/legislations reasons, data is not easily accessible. Also, with the technology rapidly growing, new problems arise frequently and in many cases it takes a while before a proper dataset is built and made publicly available. The problem investigated in this work is a clear example of the later situation. Therefore, data augmentation can be utilized not only to increase the performance of a CNN, but also to overcome the non-availability of data due to the aforementioned reasons. Applying the appropriate augmentation techniques to available datasets allows to investigate a problem where there are currently no existing public datasets and where the collection of a large training dataset poses significant challenges. Data augmentation can simulate the features and characteristics of such problems without the time, expense and data-collection and annotation challenges associated with building a dataset of many thousands of individual subjects.

In regard to our problem the main focus of the proposed augmentation techniques, is to simulate off-axis iris images as captured by a user-facing camera on AR/VR device, as to the best of our knowledge such dataset is not available. As mentioned, a possible location of the user-facing camera utilized for iris recognition and eye-gaze is below the eye. In that case the iris samples obtained will be off-axis in the horizontal and the vertical plane. Main characteristics of the iris images taken with a head-mounted device are their elliptical shape along with the fact that are not centred, in contrast with frontal iris images where the iris is most of the time centred and a more circular shape is obtained. Therefore, the augmentation techniques in this work are focused in achieving the described representation. In addition, a secondary goal of our augmentation is to introduce effects of images when captured in real-life, where the samples are not obtained in constrained conditions and a lower quality is reported. These augmentation techniques are focused on reducing the contrast between the iris and the pupil as well as adding noise to the samples.

1.4.2. Network design

In this work a low complexity network, targeted for deployment in embedded devices is designed and trained to generate the segmentation map for low quality off-axis iris images.

In order to achieve high performance results, when a network is designed, large structures with high capacity are favoured. That is translated into CNNs containing millions of parameters, which to be used require large memory and high operation cost. Therefore, executing deep CNNs requires significant hardware resources which is a limited specification in many computational platforms.

The number of parameters in the proposed network is significantly lower compared to the parameters of other deep learning approaches designed for the iris segmentation task. Thus, making the proposed network faster and with reduced memory requirements, while attaining high performance results in producing the segmentation map for off-axis iris images of low quality as represented when captured by a user-facing camera on AR/VR headset and therefore well-suited for deployment in such devices.

The rest of the paper is arranged as follows: In Section 2, the datasets used are presented along with a detailed description of the augmentation techniques. In Section 3, the network design and training are explained. In Section 4, the results are illustrated and in Section 5, the numerical evaluation of the proposed method and its comparisons with state-of-the-art segmentation techniques are presented.

Finally, it should be remarked that preliminary results from this research were first presented in Varkarakis, Bazrafkan, and Corcoran (2018). This article builds on that earlier work with more detailed and extensive experimental verifications, exhaustive description of the augmentation techniques and direct comparison on off-axis and frontal iris samples with state-of-the-art iris segmentation techniques.

2. Datasets and augmentation methodology overview

In this work, three datasets are utilized. CASIA Thousand ("CASIA Iris Image Database", 2019) and Bath800 (Rakshit, 2007) are used during the training and testing stages. UBIRIS v2 (Proenca, Filipe, Santos, Oliveira, & Alexandre, 2010) is used for tuning and testing. Two types of augmentation methods are described below. The first type is concentrated on adding real-world condition effects to the iris images, while the second is focused on augmenting the images so that they represent off-axis iris images. The combination of these two types of augmentation methods results in iris images as captured by an user-facing camera on AR/VR device. Below, the datasets used are presented along with the production of their ground truth, and finally, the augmentation techniques are explained.

2.1. Datasets

CASIA-Iris-Thousand is a subset of CASIA-Iris V4 dataset. This subset contains 20 000 iris images from 1000 subjects. The iris images are constrained, high quality and high contrast. Bath800 dataset is made of 31 997 images taken from 800 individuals. The samples similarly to the CASIA Thousand are of high quality and high contrast. Both datasets consist of Near InfraRed (NIR) samples. Finally, UBIRIS v2 dataset includes 11 102 iris images from 261 subjects, captured in visible wavelength. The samples are of low-quality as they are taken under unconstrained conditions. More detailed description of CASIA Thousand, Bath800 and UBIRIS v2 can be found in Bazrafkan et al. (2018). Samples from the datasets used in this work are shown in Figs. 2–4.

2.2. Ground truth

Bath800 and CASIA Thousand are not provided with the segmentation ground truth. However, these datasets as mentioned above contain images of high quality, high contrast and are captured under constrained conditions. In this work, the binary iris map for these datasets is produced using the commercial iris segmentation solution MIRLIN ("MIRLIN", 2019). The obtained segmentation map is considered in this work as the ground truth. The selection of the segmentation algorithm is based on the availability as well as its performance on large-scale iris evaluations (Quinn, Grother, Ngan, & Matey, 2013). The same segmentation solution was also adopted in Bazrafkan et al. (2018). The low-resolution segmentations for Bath800 and CASIA Thousand are publicly available.[1]

Regarding UBIRIS v2, the manual segmentation generated by WaveLab[2] (WaveLab, 2019), available in IRISSEG-EP dataset (Hofbauer et al., 2014), is used. The manual segmentation map is not available for all the samples of the dataset. Segmentation of only 2250 images from 50 individuals is provided and therefore only these are used in this work. Segmentations examples derived from these datasets are shown in Figs. 5–7.

2.3. Data augmentation

In order to accurately train a deep neural network, a large number of labelled training samples are required. These samples should correctly characterize the imaging problem so that it enables the deep learning process to train an accurate model. To the extent of our knowledge, there is not available a dataset with iris samples captured from a user-facing camera on AR/VR device. Even if such datasets were available it would require an accurately marked segmentation ground truth – a task which poses new problems over more conventional frontal iris images. Thus, in order to obtain a large number of labelled samples to enable the training of a DNN for AR/VR iris segmentation task, some specialized augmentations of existing datasets are required. To find the best augmentations for the iris images, precise observations have been made on iris images obtained by a user-facing camera on head-mounted displays.

The augmentation techniques are divided into two categories. The first category of augmentation techniques is focused on representing real-life scenarios where low-quality images are obtained. Based on research that has been done in Thavalengal et al. (2015b) and Thavalengal, Bigioi, and Corcoran (2015a), the difference between high-quality constrained iris images and wild ones is linked to contrast, blurring, and shadows. Consequently, to simulate the effects of real-world conditions in iris images the contrast is changed, motion blurring and shadows are added to the images. The augmentation techniques used to deteriorate the image quality and simulate unconstrained conditions are derived from Bazrafkan et al. (2018). The objective of the second category's augmentation techniques is to simulate the representation of iris images as captured by a user-facing camera on an AR/VR device. This representation includes off-axis iris images mainly of elliptical shape and not centred in the image.

The augmentation techniques are detailed in Sections 2.3.1 and 2.3.2. The workflow that is followed for the augmentation of the datasets is described in Section 2.3.3. In this work all the samples are resized to [120 × 160] using bilinear interpolation. Smaller resolution samples are preferred rather than larger ones as it accelerates the training of the deep neural network.
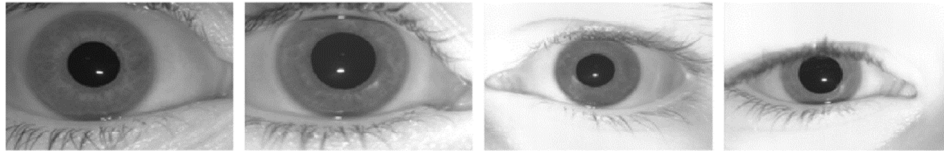
---

[1] https://Goo.gl/JVkSyG.

[2] http://www.wavelab.at/sources/Hofbauer14b/.

Fig. 2. Eye socket samples from Bath800 dataset.
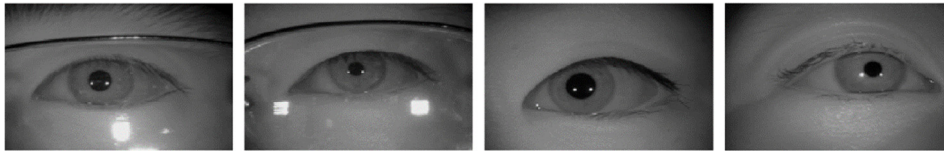
Fig. 3. Eye socket samples from CASIA Thousand dataset.
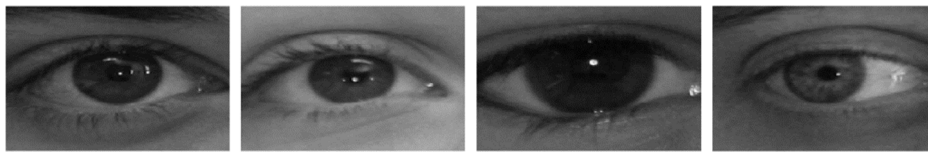
Fig. 4. Eye socket samples from UBIRIS v2 dataset.

Fig. 5. Bath800 automatic segmentation results.

Fig. 6. CASIA Thousand automatic segmentation results.

Fig. 7. UBIRIS v2 manual segmentation results.

2.3.1. Data augmentation: Simulating unconstrained conditions

The first type of augmentation techniques is applied to ensure that the samples used to train the network represent real-life scenarios. The distribution of the input data plays a vital role in what the network learns and how it will behave during the testing stage but also in unconstrained situations. As mentioned earlier, to simulate real-life captured iris images of low quality, the contrast of the samples is changed, blurring and shadows are added to the samples with the following augmentation techniques. The techniques mentioned below are derived from Bazrafkan et al.

(2018) and used with slight changes. The original code of these augmentation techniques is available.[3]

2.3.1.1. Augmentation 1: Image contrast. The iris images captured by an AR/VR device in real-world conditions compared to the high-quality, high-resolution NIR iris images acquired in constrained conditions have significant differences. The differences are with regard to the amount of contrast inside and outside

---

[3] https://github.com/C3Imaging/Deep-Learning-Techniques/blob/Iris_SegNet/DBaugmentation/DBaug.m.

Original image

Contrast reduction

Low contrast image

Iris filled mask

Fig. 8. For inside the iris region, the contrast is reduced, and the region is getting darker. The outside of iris is altered by decreasing the contrast.

Motion Blur

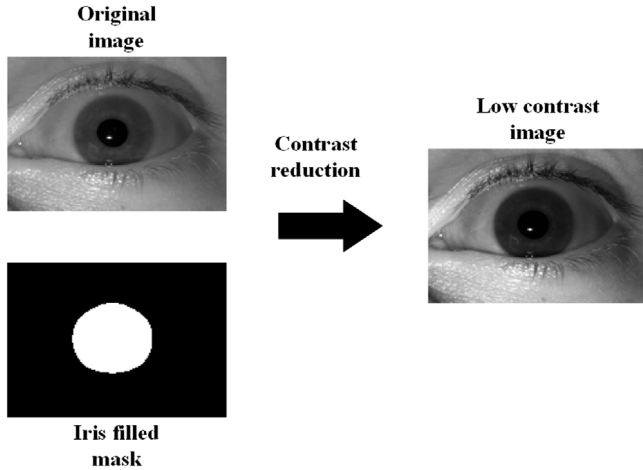Fig. 9. Applying motion blur in a random direction to the low contrast image.

Shadowing

Fig. 10. Shadowing applied to low contrast blurred image.

the iris region as in unconstrained scenarios the samples are suffering from low contrast. Another difference noted is the intensity properties of the low-quality samples inside and outside the iris region. The region inside the iris is darker than the same region in high-quality samples. For the outside region of the iris the level of brightness cannot be categorized as it could differ from overexposed and strongly bright till very dark. To bring these properties to high-quality images, the contrast inside and outside the iris region is modified separately. This is achieved with the use of histogram mapping. The following histogram mapping equations are used to reduce the contrast of the iris images. Eq. (1) is used for the region outside the iris and (2) is used for the region inside the iris.

$$y_{out} = norm(\tanh \left(3 * \left(\frac{x}{255} - 0.5 \right) \right) + \mathcal{U}(-0.2, 0.3)) * 255 \quad (1)$$

$$y_{in} = norm(\tanh \left(3 * \left(\frac{x}{255} - 0.45 \right) \right) - \mathcal{U}(0, 0.2)) * 255 \quad (2)$$

where x is the input intensity in the range [0, 255], y is the output intensity in the same range, $\mathcal{U}(a, b)$ is the Uniform distribution between a and b, and the norm function normalize the output between 0 and 1. As mentioned above the outside and inside regions of iris suffer from low contrast, but the brightness differs. For the region outside of the iris, the histogram mapping with Eq. (1) can result in bright, dark, or normally exposed low contrast outputs. For the region inside the iris, where Eq. (2) is used, the contrast is reduced while the brightness of the iris region is reduced as well. Different equations are used to reduce the contrast in the inside and outside regions of the iris so that variety is obtained. An example of this step is shown in Fig. 8.

2.3.1.2. Augmentation 2: Motion blur. Wearing AR/VR devices, head movements are inevitable. These movements can cause motion blur. Therefore, to mimic these situations and train the model in order to be efficient in these cases, motion blurring has to be introduced to the training images. In order to include this effect, the image is passed through a motion blur filter, applying the linear camera motion by $\mathcal{U}(3, 7)$ pixels in the direction $\mathcal{U}(-\pi, \pi)$, where $\mathcal{U}(a, b)$ is the Uniform distribution between a and b. The low contrast image after applying motion blur is shown in Fig. 9.

2.3.1.3. Augmentation 3: Shadowing. In unconstrained conditions, the illuminations scenarios vary. One main effect produced by different illumination directions is shadows. In order to add this effect, the iris images were multiplied with the following shadow function:
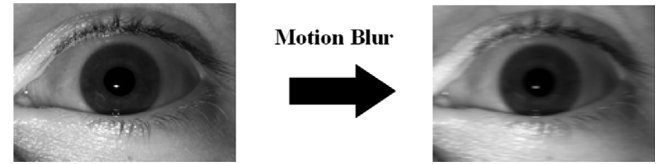
$$y = norm(\tanh(2 * randSign * (x - 0.5 + \mathcal{U}(-0.3, 0.3)))) + \mathcal{U}(0, 0.1) \quad (3)$$

where x is the dummy variable for image column number and y is the coefficient for intensity, $\mathcal{U}(a, b)$ is the Uniform distribution between a and b, the norm function normalizes the output between 0 and 1, and the *randSign* generates a random coefficient in the set $\{-1, 1\}$ which determines the direction of the shadow. The final image after applying shadowing is given in Fig. 10.

The segmentation map for these augmented samples is the same as the original segmentation ground truth as the structure and position of the iris remain unchanged.

More detailed information regarding the augmentation techniques simulating unconstrained conditions can be found in Bazrafkan et al. (2018) from where are originated.

2.3.2. Off-axis, near-perspective iris data augmentation
The second category contains two augmentation techniques, which their goal as mentioned previously is to generate iris images as they appear when acquired by a user-facing camera on AR/VR device. As noted in the introduction, a possible location of the camera used for obtaining an iris image that is to be used in iris recognition or eye-gaze is below the eye. Therefore, the iris images captured are off-axis in both horizontal and vertical planes. The augmentation techniques described below are specialized to produce such off-axis iris images. Furthermore, the combination of the two augmentation techniques and the multiple ways and different volume that each technique can be applied to an image, it allows to represent the iris regions as captured from a user-facing camera on AR/VR device from many varying perspectives, as multiple AR/VR headsets exist and each will locate the camera in a different position. This is a desired goal as it facilitates the training of generic DNN that is able to segment the off-axis iris region from user-facing cameras installed on AR/VR devices and being invariant of the camera's different set-ups. The code for these augmentation techniques is also available.[4]

2.3.2.1. Augmentation 4: Spatial stretching/contracting. The iris images when captured from an AR/VR, are characterized as distorted and with an elliptical shape. In addition, the iris is not at the centre of the image as usual. In order to generate iris images with these properties, the samples are warped by applying a spatial stretching/contracting to the iris images. The stretching is linearly applied to the images. The stretching is achieved by

---

[4] https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Off_axis_Iris.
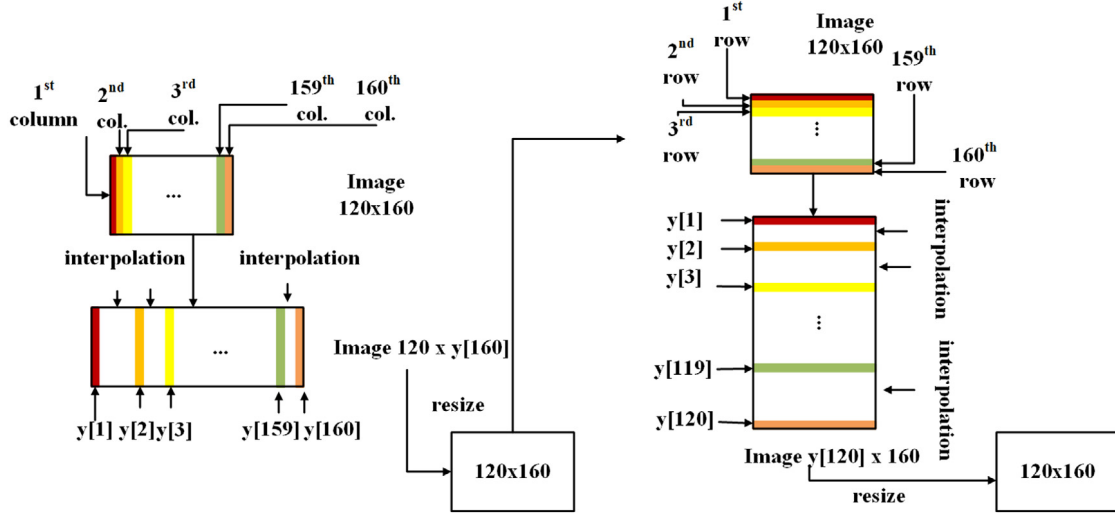
Fig. 11. Workflow of spatial stretching/contracting, illustrating the mapping of columns/rows to a new position based on y[j].

mapping every column/row of the image to a new position given from $y[j]$ as shown in Fig. 11.

The equations below illustrate how $y[j]$ is calculated for columns/rows:

$$\lambda = \mathcal{U}(2, 17) \tag{4}$$

$$k[i] = \frac{\left(\frac{1}{\lambda}\right) - \lambda}{s - 1} * t[i] + \lambda, i \in [1, s] \tag{5}$$

$$k[i] = \frac{\frac{(\lambda - 1)}{\lambda}}{s - 1} * t[i] + \lambda, i \in [1, s] \tag{6}$$

$$a = a + k[j], j \in [2, s] \tag{7}$$

$$y[j] = Round(a * 5), j \in [2, s], y[1] = 1 \tag{8}$$

where s is the length of the columns or rows of the original image depending on where the distortion is applied, $t[i]$ is a vector which includes all the integer values $[0, s - 1]$ in ascending order and $\mathcal{U}(a, b)$ is the Uniform distribution between a and b.

The first column/row of the original image is mapped on the first column/row of the stretched image. The following columns/rows are then mapped in the position determined by the value of $y[j]$. Depending on which is the desired direction for stretching the image, Eq. (5) or (6) is used in calculating $y[j]$.

The combination of (5) and (6) makes it possible to stretch the images in four main directions. If (5) is used for mapping the columns and the rows, the image will be stretched in the right and down direction. In case (6) is used for both columns and rows, the image is stretched at left and up. Using (5) when mapping the columns and the (6) when mapping the rows of the image, will result in stretching the image to the right and up direction. Finally mapping the columns using (6) and the rows using (5), the image will be stretched to the left and down direction. Each direction has the same probability of being selected when the image is stretched. For each distortion and each direction, the amount of stretching applied to the images differs on every occasion as well as the volume of the stretching applied to the columns and the rows of the image is different, so that variation is injected to the augmented dataset. The stretching is applied at first to the columns of the image. The void spaces that are created, are interpolated with a weighted nearest neighbour method, which is explained by the following equations:

$$c[i] = \frac{\frac{f(y[j])}{i - y[j]} + \frac{f(y[j+1])}{y[j+1] - i}}{\frac{1}{i - y[j]} + \frac{1}{y[j+1] - i}}, j \in [1, 160], i \in (y[j], y[j + 1]) \tag{9}$$

where $f(x)$ is a function that returns the values of the xth column/row, $c[i]$ represents the values of the ith column/row of the stretched image. The values of $y[j]$ and $y[j + 1]$ are the positions where the columns or rows of the stretched image have values and the columns/rows that need to be interpolated are located between these two positions. Finally, the image is contracted as the image is resized to the original resolution $[120 \times 160]$ using bicubic interpolation. The same process is then applied to the rows of the image. The same workflow is used for the ground truth segmentation map in order to obtain the segmentation map for the augmented sample.

The described workflow for the spatial stretching/contracting of an image is illustrated in Fig. 12. Applying spatial stretching/contracting results in an iris region that is not located in the centre of the image and with non-circular iris-pupil structures, as shown in Fig. 13 which is a usual case in iris images acquired from a user-facing camera on AR/VR headsets.

2.3.2.2. Augmentation 5: Image tilting. A possible location of the camera used for capturing the iris images, as mentioned in the introduction, will be below the eye. Therefore, the iris images should be representing samples which when captured, the camera is positioned below the eye level. To achieve that effect and also give an elliptical shape to the iris, in this second augmentation technique the samples are tilted in two directions: up and left, up and right.

A projective transformation is applied to the images. This transformation maps the top vertices of the image to a new pair of points as illustrated in Fig. 14. The values from Fig. 14, a, b, c, and d are randomly generated between a range of values, so the image is tilted in the desired direction with variation. When the image is tilted up and left the values of a, b are in $\mathcal{U}(0.15, 0.45)$, c in $\mathcal{U}(0.9, 1)$ and d in $\mathcal{U}(0, 0.1)$. When the image is tilted up and right, the values of a, b are in $\mathcal{U}(0, 0.1)$, c is in $\mathcal{U}(0.55, 1)$ and d is in $\mathcal{U}(0.15, 0.45)$, where $\mathcal{U}(a, b)$ represents the Uniform distribution between a and b. During this transformation as the image shrinks the interpolation used is the nearest-neighbour. The probability of the images being tilted in a direction between the two options (up and left/up and right) is the same.

As shown in Fig. 14, when the transformation is applied, while mapping the top vertices, to the a, b, c and d points, the image is compressed at the boundaries. Since the resolution of the image has to stay unchanged, the void spaces around boundaries should be filled to avoid sharp edges in the image. Since the void spaces are at the boundaries of the image, there is not a direct way
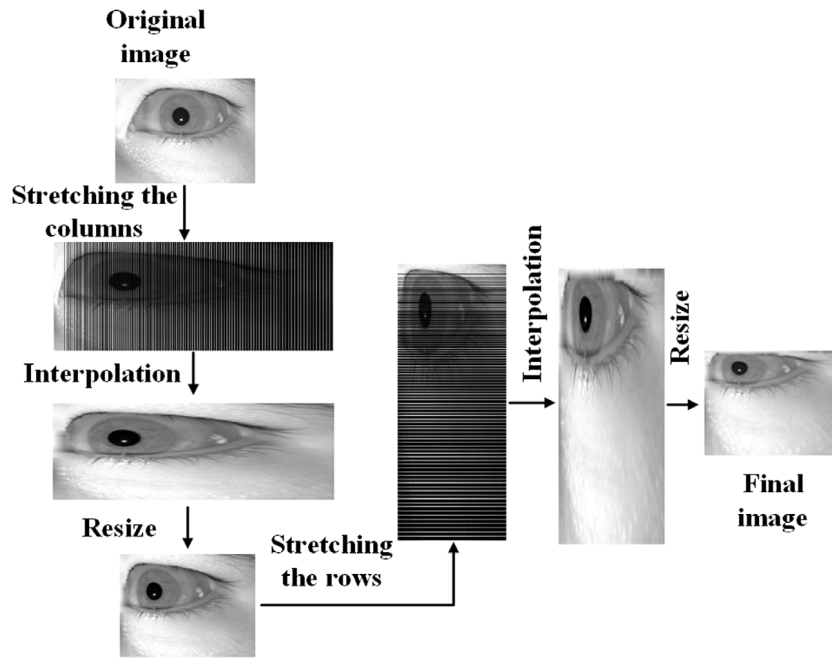
Fig. 12. Workflow of spatial stretching/contracting. For this transformation Eqs. (6) was used to map the columns and the rows of image and direct the image in up and left direction.
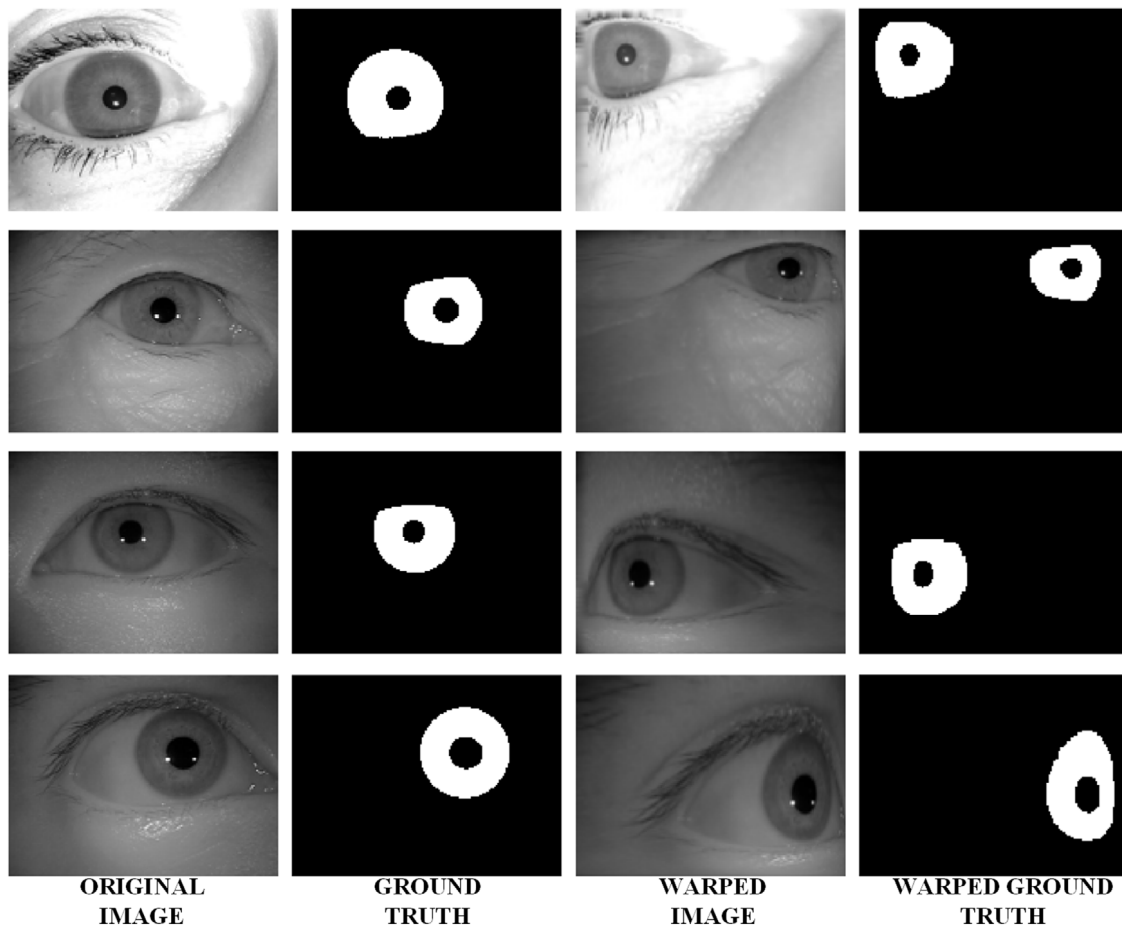

| ORIGINAL IMAGE | GROUND TRUTH | WARPED IMAGE | WARPED GROUND TRUTH |

Fig. 13. Spatial stretched/contracted (warped) samples and their corresponding segmentation map.
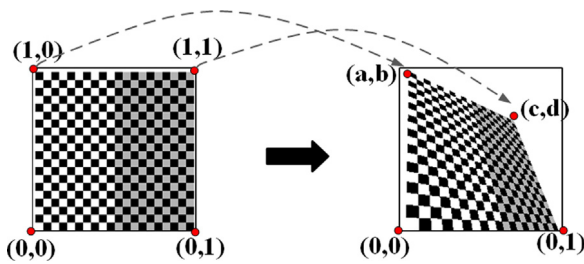
Fig. 14. Tilt transformation.

to apply interpolation. Therefore, the value from the edges of the tilted image is extended for each column up to the image boundary. The same process is applied to the image rows. After this process has been applied in both columns and rows, the average value is assigned to the void spaces. Finally, in order to smooth the interpolated areas of the image, a gaussian 3×3 filter with standard deviation σ equal to 2 is applied to this region.

The described workflow for tilting an image is shown in Fig. 15, and in Fig. 16 samples are shown where the tilting transformation is applied. To obtain the segmentation map for the augmented samples, the same workflow is applied to the segmentation ground truth with the only difference being that the void spaces created are filled with black.

2.3.3. Dataset preparation
2.3.3.1. Workflow of combining the augmentations techniques. The augmentation techniques are combined in various ways so that the dataset represents a generalized and realistic scenario and as a result the trained model can be robust and perform well in all the different conditions that one can encounter with iris images acquired from a user-facing camera on AR/VR device.

The augmentation techniques are mixed in three ways. Samples are augmented by only using the methods for simulating the off-axis, near-perspective iris images. At first, the spatial stretching/contracting transformation is applied to an image with 50% probability. In the next step, the tilting transformation is applied to the rest of the samples that the first transformation was not applied to. In addition to these, for an image that spatial stretching/contracting is applied to in the first step, there is a 50% probability that tilting is applied afterwards. In the second way, the samples are augmented using only the methods for simulating unconstrained conditions. At first, the contrast of all the iris images is modified as explained. Afterwards, all the images are passed through the motion filter, and finally, the technique used to introduce shadows is applied to the image. The probability that shadows are added to an image is 50%. Thirdly, the techniques from the two augmentation categories are combined. Initially, the techniques simulating the off-axis, near perspective iris images are applied to an image based on the augmentation workflow described above. Later the same image is processed using the techniques simulating unconstrained conditions, including contrast reduction, motion blurring, and shadowing. In Fig. 17, the workflow explained is illustrated.

The augmented samples simulate iris images captured using a user-facing camera on AR/VR device, frontal iris images affected by unconstrained conditions and AR/VR images affected by unconstrained conditions. Bath800 and CASIA Thousand are augmented with all three combinations of the augmentation techniques described. UBIRIS v2 was augmented only by using the augmentation techniques simulating iris images as represented by an AR/VR device. The samples of this dataset as mentioned previously are captured in unconstrained conditions, and therefore it will be redundant to make use of the augmentation techniques that simulate real-world conditions as they already exist in the dataset of UBIRIS v2.

2.3.3.2. Dataset analysis. In this section a further analysis of the workflow used to combine the augmentation techniques is presented, in order to provide a better insight of the dataset created and used in this work.

As mentioned above, the workflow was designed in that way so that the dataset created to train the network represents a generalized and realistic problem. By using the three combinations of the augmentation techniques, three different subsets are created as shown in Fig. 17, that form the main dataset used in this work. The first combination as described earlier uses only the augmentation techniques designed to simulate off-axis iris images. This process is used twice for each dataset creating the off-axis iris subset. The second combination uses only the augmentation techniques that simulate unconstrained conditions. This process is used once for each dataset consisting thus the unconstrained condition subset. Finally, the third combination, uses the augmentation techniques simulating the off-axis iris images and unconstrained conditions. This process is used twice for each dataset formulating the off-axis & unconstrained condition iris subset. Bath800 and CASIA Thousand combined consist of around 50.000 samples. With the use of the described workflow, 250.000 augmented samples are created. The off-axis iris subset is 100.000 samples, the unconstrained condition subset is 50.000 samples and 100.000 more samples from the unconstrained condition and off-axis iris subset. With the addition of the 50.000 original samples from Bath800 and CASIA Thousand, the final dataset used consists of 300.000 samples. In Table 1, a further analysis is presented describing the percentage of samples, with each augmentation technique or their combination, to the dataset.

Regarding the UBIRIS v2 as stated above, it consists of samples acquired in unconstrained conditions and therefore there is not a necessity of augmenting the samples with the augmentation techniques simulating unconstrained conditions. The samples of UBIRIS v2 are augmented only with the use of the augmentation techniques simulating off-axis iris images. This procedure is operated twice, creating the off-axis iris subset of UBIRIS v2 which along with the original samples are used in this work.

Finally, the element of randomness introduced in the augmentation techniques as well as at the way they are combined as explained in the workflow plays an important role to the augmentation process. One instance of that is that the direction of the shadowing, stretching or tilting is chosen randomly for each image. Also, the volume that an augmentation technique is applied to an image is chosen randomly between a range of values. Additionally, as illustrated in the workflow a sample is augmented with one or more augmentation techniques combined in different ways. These three approaches make it possible that a variety of conditions are introduced into the dataset leading to a generalized solution and each time producing unique samples with different characteristics and distributions. Examples of augmented samples with the use of all three different workflow combinations and their corresponding ground truth are given in Fig. 18.

3. Network design & training

In this section the design of the network is presented along with a detailed comparison of its complexity with other CNN methods designed for the iris segmentation task followed by the procedure of training and fine-tuning.

3.1. Network design

For the segmentation task, a fully convolutional network inspired by Bazrafkan et al. (2018) is used, consisting of 10 layers. The network starts with a 3×3 kernel mapping the input (1
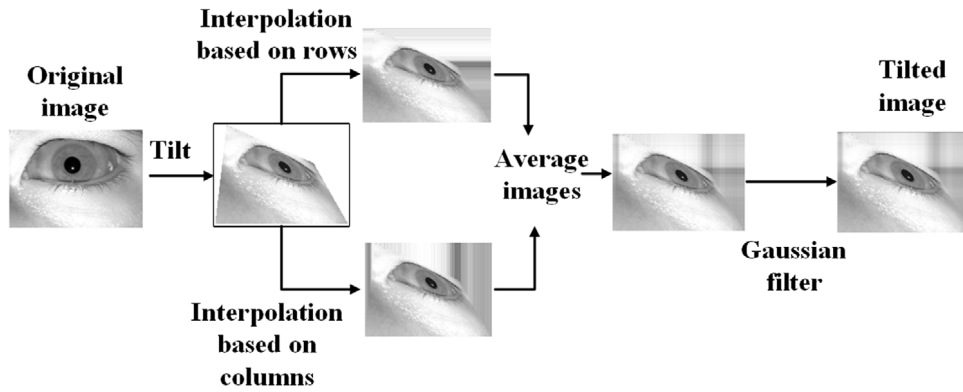
Fig. 15. Workflow of image tilting.

| ORIGINAL IMAGE | GROUND TRUTH | TILTED IMAGE | TILTED GROUND TRUTH |
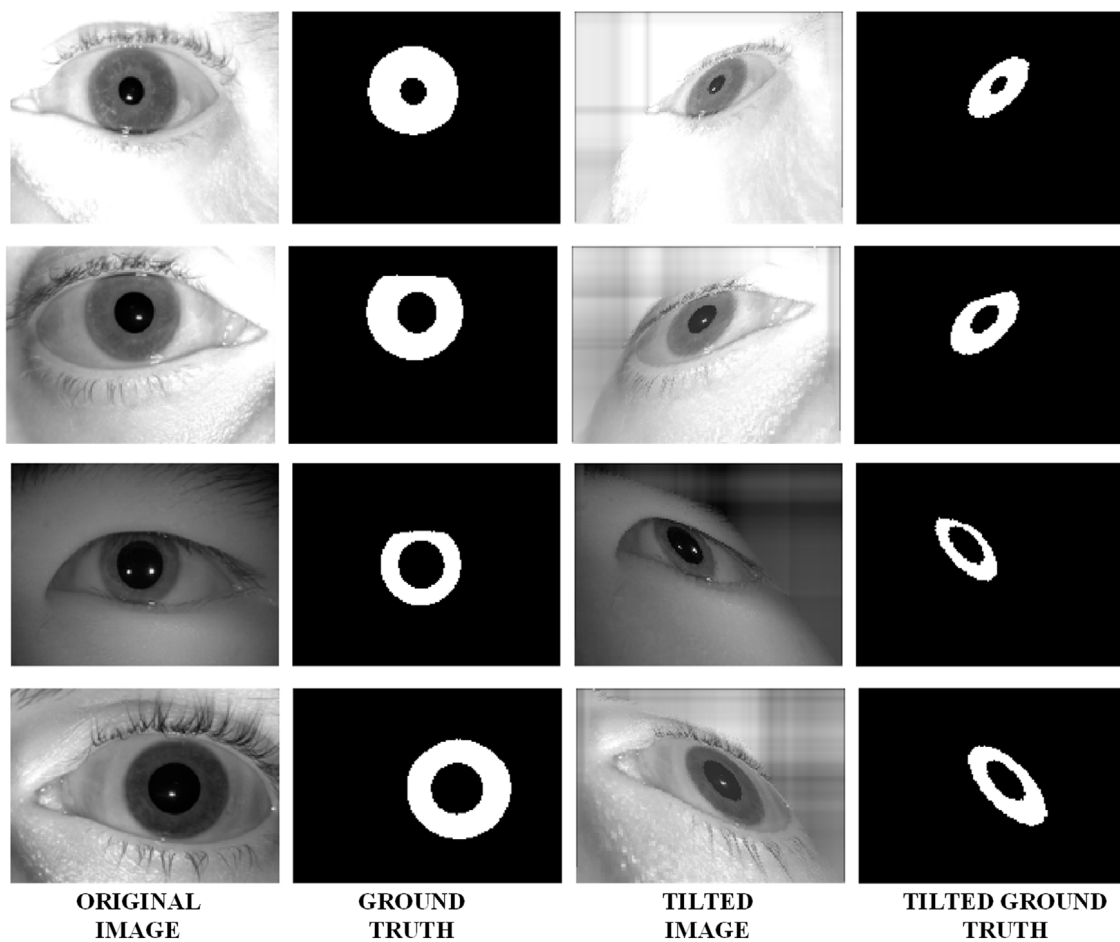
Fig. 16. Tilted samples and their corresponding segmentation map.

channel) on the first convolutional hidden layer which consists of 32 channels using a rectified linear unit (ReLu) as an activation function. The kernel size remains the same throughout the hidden convolutional layers, as well as, the number of channels and their activation function. Finally, at the output layer (1 channel), the kernel size is 3 × 3, but in this layer, the sigmoid activation function is used. Pooling layers were not used as it was observed that the performance of the network's output was decreasing. The design of the network is illustrated in Fig. 19.

3.2. Complexity comparison of CNNs for iris segmentation

In this section, the complexity of several CNNs for iris segmentation will be compared with the proposed method. When referred to the complexity of a CNN, the main characteristics that one shall investigate is the number of parameters, the memory requirements for storing the parameters and the number of multiply-accumulate operations (MAC).

Fig. 17. Workflow of augmentation techniques.

Fig. 18. Augmented samples and their corresponding ground truth.

Table 1

Percentage (%) of images with each augmentation technique or combination in the dataset. In this table the augmentation techniques are referred as Contrast reduction: Contrast, Motion blur: Blur, Shadows: Shadows, Spatial stretching/contracting: Warp and Tilting: Tilt.

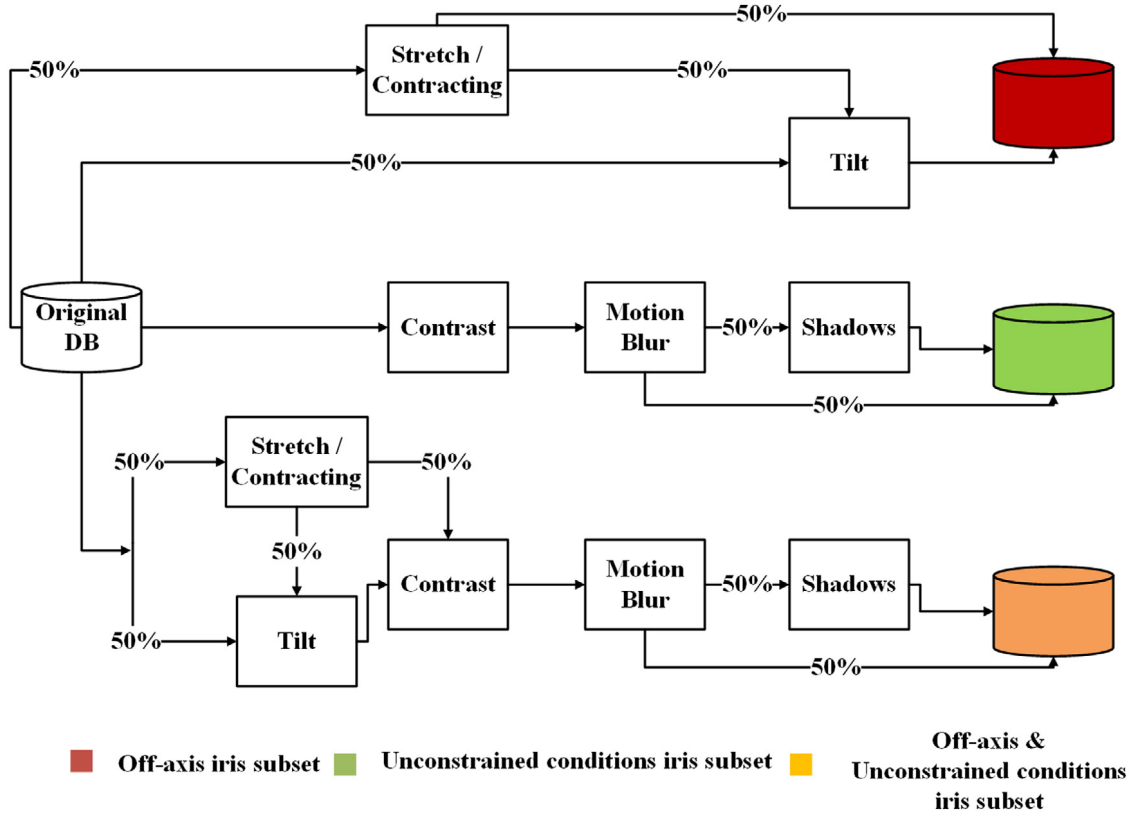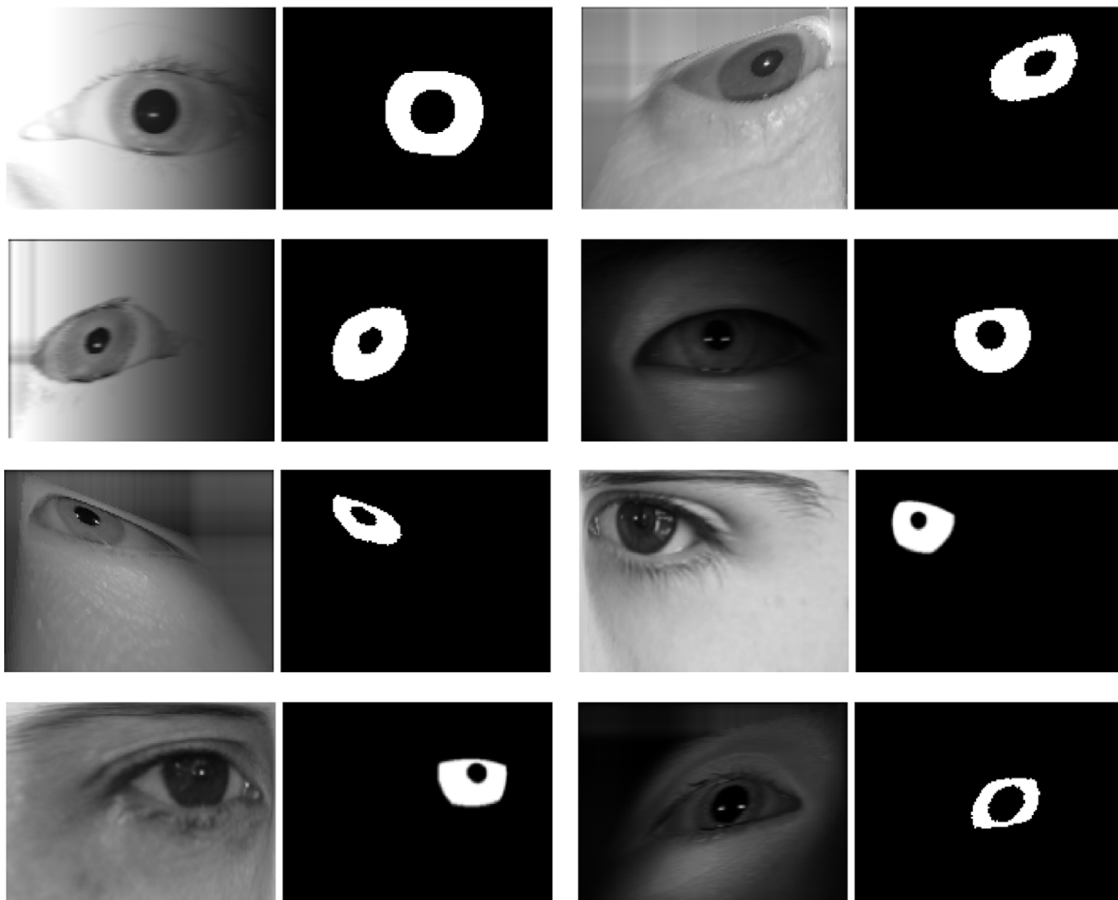| Augmentation techniques	% of images	Dataset
Contrast & Blur	~8.5%	Unconstrained condition subset
Contrast & Blur & Shadows	~8.5%	
Warp	~8.5%	Off-axis subset
Tilt	~16.5%	
Warp & Tilt	~8.5%	
Warp & Contrast & Blur	~4%	Off-axis
Tilt & Contrast & Blur	~8.5%	&
Tilt & Contrast & Blur & Shadows	~8.5%	Unconstrained conditions subset
Warp & Contrast & Blur & Shadows	~4%	
Warp & Tilt & Contrast & Blur	~4%	
Warp & Tilt & Contrast & Blur & Shadows	~4%	
No augmentation	~16.5%	Original subset
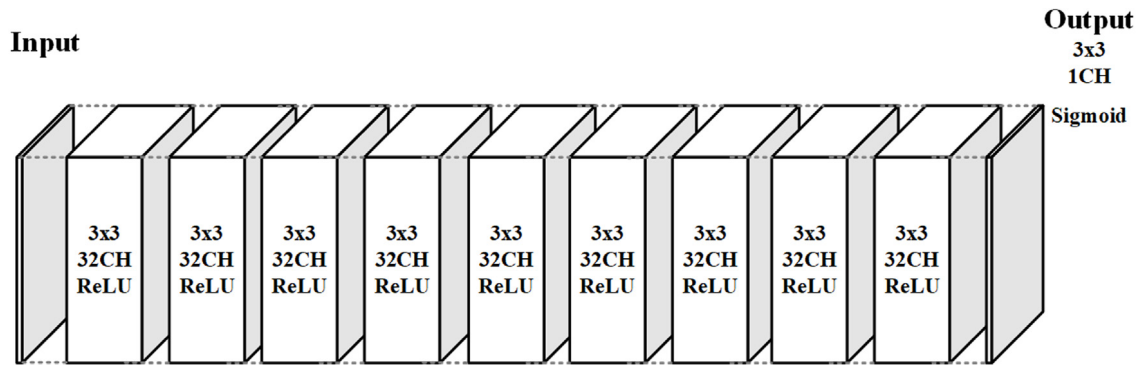
Fig. 19. Network Design.

It is common practice that when an architecture of CNN is designed, that deep and large structures are favoured thus increasing the possibility of solving the investigated problem or promise higher performance from a smaller size CNN. Selecting a CNN with a deeper structure rather than a more compact structure, comes with some drawbacks such as increased training and execution time as well as generous memory requirements. There are cases, such as the proposed CNN, where a low complexity network can produce similar results as a high complexity network and as extension make it feasible to eliminate the downsides of a large CNN.

The proposed CNN consists of less than 75k parameters, requiring only 0.28 MB of memory to store the parameters and 1426.64M MAC for an input image with dimensions [120 × 160 × 1] [*width × height × channels*]. The SPDNN (Bazrafkan et al., 2018) consist of more than 1M parameters, requiring 35.26 MB to store them and 13536.22M MAC for a smaller input image of dimensions [98 × 128 × 1]. Another high performance deep learning method, MFCNs (Liu et al., 2016) is of high complexity and memory requirements, with 21M parameters, needing 82.56 MB of memory to store them. The input dimension and MAC in this structure are not specified as the input image dimension is not fixed and the number of MAC is related to the dimension of the input image. The complexity characteristics of the methods mentioned are shown in Table 2.

In this section is presented the low complexity proposed network, with reduced memory requirements resulting into a more efficient solution which is compatible for deployment in embedded applications such as AR/VR headsets. Furthermore, in the evaluation of the proposed network in Section 5, is demonstrated that the low complexity network proposed in this work can obtain high performance iris segmentation results in both off-axis and frontal samples. The proposed CNN is outperforming

Table 2

Complexity of CNNs for iris segmentation.

Metrics	Methods		
	Proposed method	SPDNN	MFCNs
Total no. parameters	74.593	1.101.851	21.643.596
Parameters size	0.28 MB	35.26 MB	82.56 MB
Input size dimensions (*width × height × channels*)	120 × 160 × 1	96 × 128 × 1	N/A
Total MAC	1426.64M	13536.22M	N/A

other methods in segmenting off-axis iris images. Also, despite the fact that the network is designed for segmenting off-axis iris images, the results reported in segmenting frontal iris images are comparable to the-state-of-the-art SPDNN method of higher complexity and memory requirements.

3.3. Training and fine-tuning

3.3.1. Training

The network is trained on the original and augmented samples of Bath800 and CASIA Thousand. The dataset is divided 70% for the training set, 20% for validation set and 10% for the test set.

The training was carried out in TensorFlow library. The Mean Squared Error is used as the loss function. The Gradient Descent with Adaptive Moment Estimation (Adam) is used, with a learning rate of 1e–4, beta1 and beta2 equal to 0.9 and 0.999 respectively, to optimize the loss function. The training is done on a desktop computer with Nvidia GTX 1080 GPU. The executable of the trained network is available at[7].

3.3.2. Fine-tuning

In this section the process of fine-tuning the original network with the UBIRIS v2 dataset is described. Fine-tuning is a concept of transfer learning. Transfer learning is a machine learning technique, where knowledge gain during training in one type of problem is used to train in another related task or domain.

The proposed model was trained on the augmented and original samples from Bath800 and CASIA Thousand. UBIRIS v2 differs from the other datasets in the fact that it consists of visible iris image while Bath800 and CASIA Thousand are taken in NIR domain. Obtaining high-performance segmentation results in visible iris samples requires training a new model from the beginning or either fine-tune a pre-trained network on a dataset with visible samples. As UBIRIS v2 is a small dataset, training a new model is not possible, therefore fine-tuning the parameters of the pre-trained network is more functional. The network is trained on NIR iris samples and therefore it is excepted that the network transfers the information and tune the parameters on the UBIRIS v2 samples, as the context of the task and the datasets are similar.

Regarding the specifics of fine-tuning, the network is fine-tuned on the augmented and original samples of UBIRIS v2. The dataset is divided 70% for the training set, 20% for validation set and 10% for the test set. The training was carried out in TensorFlow library. The Mean Squared Error is used as the loss function. The Gradient Descent with Adaptive Moment Estimation (Adam) is used, with a learning rate of 5e-5, beta1 and beta2 equal to 0.9 and 0.999 respectively, to optimize the loss function. The tuning is done on a desktop computer with Nvidia GTX 1080 GPU.

4. Results

The input of the network is a greyscale iris image of 1 channel with dimensions [120 × 160]. The output of the network is a greyscale segmentation map with values between [0, 1] and of the same size and channels as the input. The binary segmentation map is obtained by using a thresholding technique, where the values bigger than the threshold are shifted to 1 and the others to 0. The threshold value 0.55 is used in this work for the Bath800, CASIA Thousand which are datasets containing NIR images. Regarding UBIRIS v2 which contains visible samples, after fine-tuning the network to the dataset, the threshold with value 0.4 is selected. The output of the proposed model for the different datasets is shown in Figs. 20–22.

5. Evaluation

Several metrics are used to evaluate the proposed method and conduct a detailed comparison with several segmentation methods of the literature. The metrics used in this work are: accuracy, sensitivity, specificity, precision, NPV and F1-score. More information about these metrics can be found in Bazrafkan et al. (2018). Two main experiments have been used to evaluate the performance of the proposed network:

(1) Evaluate the proposed network on the off-axis augmented samples:
The network is trained on the original and augmented samples of Bath800 and CASIA Thousand. The network is tested on the off-axis augmented samples of Bath800, CASIA Thousand and UBIRIS v2. These are the off-axis subset and off-axis with unconstrained condition subset for Bath800 and CASIA Thousand and the off-axis subset for UBIRIS v2, as described in Section 2.3.3. In continuance it is compared with the segmentation results on these samples from the methods: SPDNN (Bazrafkan et al., 2018), IrisSeg

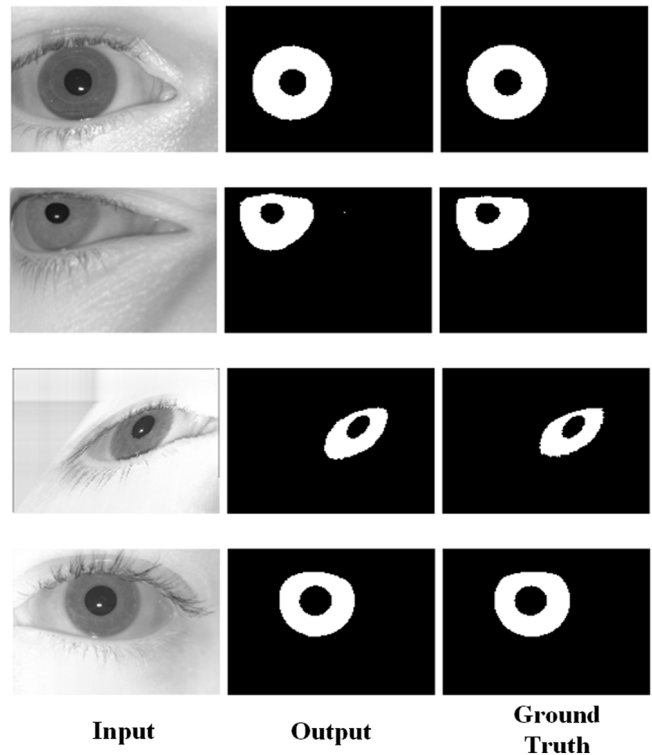**Fig. 20.** Output of the network for the augmented off-axis and original samples of Bath800.
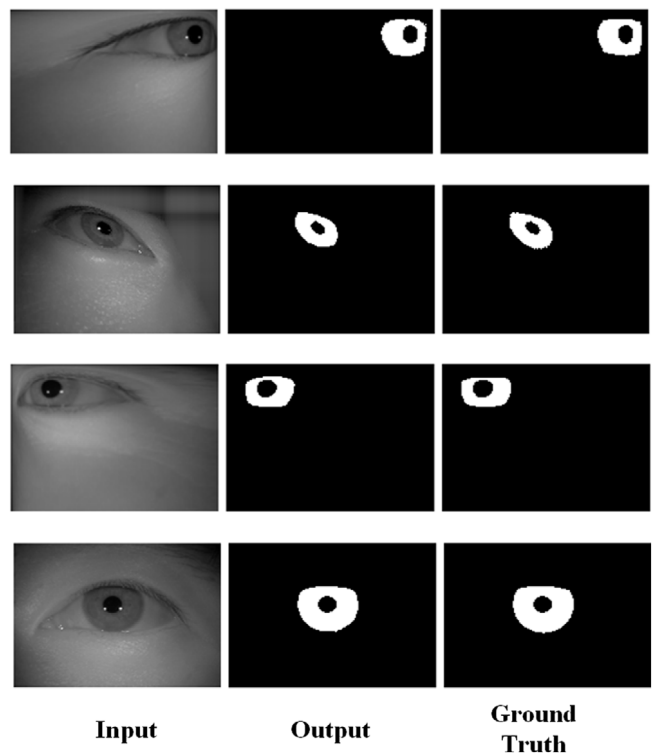

Fig. 21. Output of the network for the augmented off-axis and original samples of CASIA Thousand.

(Gangwar, Joshi, Singh, Alonso-Fernandez, & Bigun, 2016) and OSIRIS (Othman, Dorizzi, & Garcia-Salicetti, 2016). The test set of the augmented samples is used to test the network and the other methods.
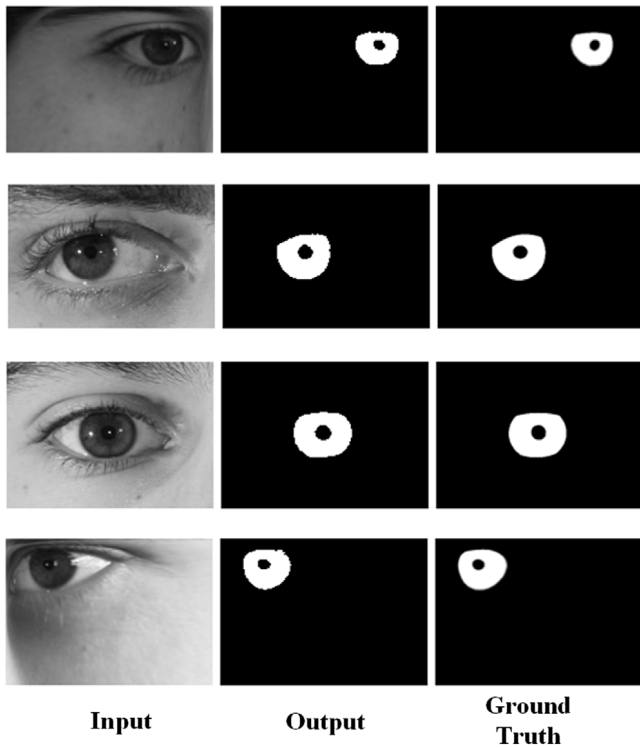
Fig. 22. Output of the network for the augmented off-axis and original samples of UBIRIS v2.

(2) Evaluate the network on the original samples from the datasets:

The network is tested on the original samples of Bath800, CASIA Thousand and UBIRIS v2, which consist of frontal iris samples. The test set of these datasets is used for testing the proposed method. The results of the proposed network are compared extensively with the state-of-the-art SPDNN on the Bath800, CASIA Thousand and UBIRIS v2. Furthermore, the results of the network are compared with the available results from other segmentation methods of the literature.

The results presented on UBIRIS v2 are the results of the original network after tuning.

5.1. Evaluation and comparison on off-axis augmented samples

In this section the proposed method is tested on the off-axis augmented samples. These samples are the combination of the off-axis subset and the off-axis with unconstrained condition subset for Bath800 and CASIA Thousand and the off-axis iris subset for UBIRIS v2. The test sets from the datasets are used for the testing stage.

5.1.1. Evaluation

The proposed network produces high performance results in the datasets Bath800 and CASIA Thousand. This is expected as the network is trained on them. On UBIRIS v2 the network is able to provide accurate segmentation results but is not able to perform at the same level as on Bath800 and CASIA Thousand. The samples of UBIRIS v2 are taken in visible spectrum and therefore the distribution differs. The proposed network with tuning is able to produce high segmentation results showing that the CNN is able to adopt to a similar task but with different distribution.

Table 3

Comparison of the proposed method with other segmentation methods on the off-axis augmented samples of Bath800. A higher value for μ and lower for σ is desired.

Metrics		Bath800			
		Proposed method	SPDNN	IrisSeg	OSIRIS
Accuracy	μ	99.22%	97.03%	96.10%	95.86%
	σ	0.62%	1.96%	3.53%	2.80%
Sensitivity	μ	92.98%	58.71%	67.26%	62.16%
	σ	8.7%	38.04%	21.82%	35.72%
Specificity	μ	99.62%	99.15%	98.02%	98.00%
	σ	0.38%	0.86%	3.53%	2.37%
Precision	μ	93.97	80.34%	75.88%	67.68%
	σ	7.41%	19.32%	21.18%	24.11%
NPV	μ	99.52%	97.74%	97.79%	97.60%
	σ	0.57%	2.14%	1.72%	2.24%
F1-score	μ	93.21%	59.90%	68.63%	59.54%
	σ	7.70%	35.76%	19.51%	31.78%

In the datasets that the network is trained on, the accuracy and the F1-score and sensitivity measurements are higher showing high quality in returning true results and more consistent segmentations in comparison with UBIRIS v2. The same applies with the sensitivity and NPV metrics showing that the network is able to rule-out non-iris pixels more effectively in the trained datasets than the dataset that it was tuned on. Although, the network has higher performance in precision and specificity on the UBIRIS v2 dataset, showing greater capability in returning real iris pixels in the UBIRIS v2 dataset rather than the Bath800 and CASIA Thousand. The results are shown in Tables 3–5.

5.1.2. Comparison with SPDNN, IrisSeg and OSIRIS

The proposed method is designed for segmenting low quality off-axis iris images as acquired from an AR/VR device. The proposed method is compared with the SPDNN, IrisSeg and OSIRIS on the test set of the augmented off-axis samples. The selection of these algorithms is based on their availability. Furthermore, the SPDNN is a state-of-the-art segmentation method specialized on low quality iris images and IrisSeg and OSIRIS are well-established methods with high performance in the iris segmentation task. The SPDNN is trained on the original and augmented samples of Bath800 and CASIA Thousand and tuned on UBIRIS v2. The augmented samples used in their work are representing unconstrained scenarios. The SPDNN is a network with high capacity and large number of parameters as analysed earlier.

The SPDNN when tested on the off-axis augmented samples is able to provide overall good results in accuracy and specificity and average results in precision. The performance of the SPDNN is low in the sensitivity and F1-score measurements. The proposed network is outperforming the SPDNN in all the evaluation metrics showing higher results and ability to segment off-axis iris samples as appear when acquired from a user-facing camera on AR/VR device. In regard to IrisSeg and OSIRIS there not able to provide high segmentation results for the augmented off-axis samples. The low performance results of IrisSeg and OSIRIS are due to the fact that the augmented samples that used are challenging as they simulate off-axis iris images in unconstrained conditions. In addition, IrisSeg and OSIRIS were not able to provide a segmentation in many cases. The results included for IrisSeg and OSIRIS are only for the images that the algorithms were able to provide a segmentation. The results are given in Tables 3–5.

Table 4
Comparison of the proposed method with other segmentation methods on the off-axis augmented samples of CASIA Thousand. A higher value for μ and lower for σ is desired.

Metrics		CASIA Thousand			
		Proposed method	SPDNN	IrisSeg	OSIRIS
Accuracy	μ	99.40%	97.75%	96.7%	95.81%
	σ	0.56%	1.66%	5.52%	2.49%
Sensitivity	μ	90.64%	49.36%	69.67%	36.34%
	σ	11.14%	43.15%	27.13%	38.40%
Specificity	μ	99.77%	99.43%	97.90%	98.60%
	σ	0.29%	0.9%	5.62%	2.13%
Precision	μ	94.17%	75.89%	74.37%	48.42%
	σ	7.87%	28.26%	27.97%	34.48%
NPV	μ	99.59%	98.27%	98.63%	97.07%
	σ	0.49%	1.64%	1.19%	2.14%
F1-score	μ	91.93%	49.40%	69.00%	35.83%
	σ	9.66%	41.44%	26.72%	34.31%

Table 5
Comparison of the proposed method with other segmentation methods on the off-axis augmented samples of UBIRIS v2. A higher value for μ and lower for σ is desired.

Metrics		UBIRIS v2			
		Proposed method	SPDNN	IrisSeg	OSIRIS
Accuracy	μ	98.83%	97.94%	87.17%	92.96%
	σ	1.16%	1.84%	9.10%	5.31%
Sensitivity	μ	83.89%	60.17%	27.06%	24.11%
	σ	10.48%	34.20%	23.51%	29.04%
Specificity	μ	99.77%	99.75%	91.03%	97.58%
	σ	0.46%	0.73%	9.30%	4.15%
Precision	μ	95.26%	93.78%	24.31%	39.11%
	σ	9.87%	15.51%	29.21%	38.34%
NPV	μ	98.94%	98.01%	94.97%	95.15%
	σ	1.12%	1.88%%	4.22%	4.33%
F1-score	μ	88.72%	66.35%	21.58%	23.58%
	σ	10.62	35.49%	22.50%	29.66%

5.2. Evaluation and comparison on the frontal iris-region samples

In this section the proposed method is evaluated and compared on the frontal original samples of Bath800, CASIA Thousand and UBIRIS v2, which consist of frontal iris samples. It is worthwhile to note that the proposed technique is designed for segmenting off-axis consumer level iris images. Despite that, the experiments below are carried out in order to conduct a fair comparison with the other methods on frontal images. Meanwhile the proposed method is giving the best results on segmenting the augmented off-axis samples.

5.2.1. Evaluation on the frontal iris-region samples

The proposed network is now tested on the original samples from Bath800, CASIA Thousand and UBIRIS v2. For this procedure the test sets of the datasets are used.

Similar outcomes with the one's on the evaluation of the proposed method on the off-axis iris samples are found in the evaluation of the original samples. The proposed network has higher performance in the datasets that the network is trained on, Bath800 and CASIA Thousand. Lower performance is reported on UBIRIS v2. The network accomplishes high accuracy results in all datasets showing that has high quality in returning true results. Moreover, in all datasets it returns high values in specificity and precision, meaning that the model performs well returning iris pixels. The sensitivity metric on Bath800 and CASIA Thousand is

Table 6
Testing the proposed method on the original samples of several datasets.

Metrics		Proposed method		
		Bath800	*CASIA Thousand*	*UBIRIS v2*
Accuracy	μ	99.13%	99.50%	98.92%
	σ	0.52%	0.36%	0.67%
Sensitivity	μ	94.90%	94.67%	88.38%
	σ	6%	4.33%	9.29%
Specificity	μ	99.56%	99.86%	99.71%
	σ	0.47%	0.16%	0.39%
Precision	μ	95.67%	97.39%	96.33%
	σ	6.33%	2.83%	7.22%
NPV	μ	99.49%	99.63%	99.10%
	σ	0.45%	0.34%	0.60%
F1-score	μ	95.17%	95.94%	91.46%
	σ	5.43%	2.89%	9.63%

high, showing the ability of the model in ruling out non-iris pixels accurately while in UBIRIS v2 the same metric has average performance. The same applies to the F1-score measurement showing that the network produces more consistent segmentations, both in finding iris and non-iris pixels in the datasets Bath800 and CASIA Thousand compared to UBIRIS v2. In Table 6 the results of the proposed network on the test sets of the original samples from Bath800, CASIA Thousand and UBIRIS v2 are presented.

5.2.2. Comparison with the SPDNN

The SPDNN is a sophisticated network, with state-of-the-art results in the iris segmentation task. Now as mentioned earlier the SPDNN was trained on samples of Bath800 and CASIA Thousand and tuned on the UBIRIS v2, as is the proposed method. The SPDNN is of high complexity with 14 times more number of parameters when compared to the proposed network. This is an aspect that should be considered in the comparison between these segmentation methods. Also, as mentioned earlier the proposed network is designed for segmenting off-axis iris images as captured by a user-facing camera on AR/VR device. The numerical results of the SPDNN (Bazrafkan et al., 2018) performance are reported as presented in their work.

5.2.2.1. Comparing results on Bath800, CASIA Thousand and UBIRIS v2.
The proposed method shows higher accuracy than the SPDNN in the Bath800 dataset which implies better quality in returning true results. The performance in specificity of the proposed method is also higher than the SPDNN. However in the precision metric the SPDNN is performing better. That shows that both can perform well in returning iris pixels, with not one method being better than the other. The same applies in the ability of the methods in ruling out non-iris pixels, as in NPV the proposed method is performing better than the SPDNN while the SPDNN shows higher results from the proposed method in the sensitivity metric. On the other hand, a small advantage of the SPDNN over the proposed method is in the F1-score showing a better efficiency. Overall in Bath800 dataset there is not a clear advantage of one method over the other as the performance in the metrics is divided with the differences between them either in favour or against them being marginal. The proposed network is performing comparable with the SPDNN in the Bath800 dataset.

In regard with the CASIA Thousand dataset, the SPDNN shows a small advantage over the proposed method. The proposed method performs better in the specificity and precision metrics showing higher quality in returning iris pixels than the SPDNN. In the rest of the evaluation metrics the SPDNN is performing better than the proposed method. Nonetheless, generally the differences in performance are marginal.
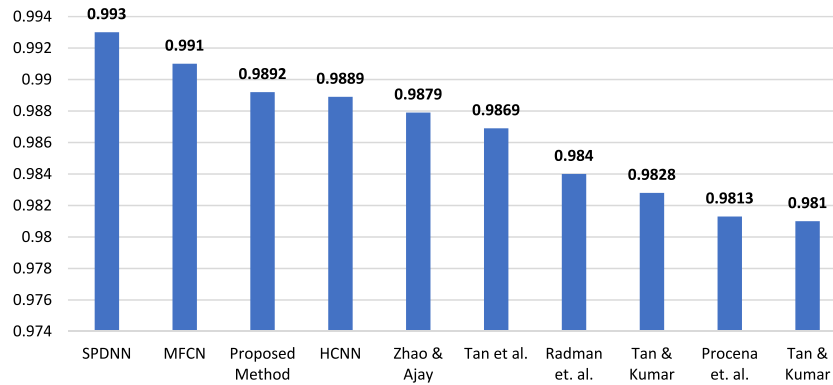
Fig. 23. Accuracy of the proposed method vs. other methods over the original UBIRIS v2 dataset.

On UBIRIS v2, the SPDNN performs better than the proposed method. The proposed method is performing better only in the specificity and precision metrics showing that is better on returning iris pixels than the SPDNN. In some metrics such as accuracy and NPV the difference is marginal showing that the proposed method is almost as good as the SPDNN in returning true results and in ruling out non-iris pixels. In the rest of the metrics there is a slight difference between the two methods, showing that the SPDNN is able to adopt better to the dataset that the methods are tuned utilizing thus the larger number of parameters of the SPDNN.

Overall, the proposed network and SPDNN performs similarly in the Bath800 and CASIA Thousand datasets, which are the datasets that were trained on. Therefore, showing that when trained the proposed network is comparable to the SPDNN despite that the complexity of the proposed network is at least an order of magnitude less than the SPDNN as analysed earlier. In the UBIRIS v2, where the proposed network and the SPDNN are tuned, the proposed method shows high results but the SPDNN still outperforms it, showing the ability to adopt better to a different dataset distribution, utilizing the higher complexity of its structure. The comparison between the two methods is shown in Table 7.

5.2.3. Comparison with state-of-the-art methods

In the following section the proposed method is compared to the most advance and state-of-the-art segmentation methods in the literature. First, accuracy over the challenging UBIRIS v2 dataset is compared with several methods. In continuance, it is evaluated and compared over CASIA Thousand and UBIRIS v2 in three important segmentation metrics: sensitivity, precision and F1-score with known segmentation methods.

5.2.3.1. Comparison of accuracy on UBIRIS v2. The accuracy of the proposed method is compared with state-of-the-art segmentation methods over UBIRIS v2. The state-of-the-art segmentation methods used in the comparison are the SPDNN (Bazrafkan et al., 2018), MFCN and HCNN from Liu et al. (2016) and Total Variation (TV) model utilized in Zhao and Ajay (2015); also an integrodifferential constellation followed by a curvature fitting model proposed in Tan, He, and Sun (2010), the HOG-SVM from Radman, Zainal, and Suandi (2017), and the random walker algorithm used in Tan and Kumar (2013). Moreover, the method proposed in Proenca (2010) where the sclera and iris regions are detected separately using neural networks as classifiers, and polynomial fitting is applied estimating the final iris region and finally the method from Tan and Kumar (2012) in which proposes a post-classification procedure including reflection and shadow removal and several refinements on pupil and eyelid localizations to get higher performance on iris segmentation task. The

accuracy of the proposed method compared with the aforementioned state-of-the-art methods on UBIRIS v2 dataset is presented in Fig. 23.

As illustrated, the proposed method has the third best performance compared with the state-of-the-art segmentation methods. The two methods that are performing better are: the SPDNN of Bazrafkan et al. (2018) and MFCN of Liu et al. (2016). However, these two methods are considerably more complex. As analysed in Section 3.2, SPDNN and MFCN contain 1M and 21M parameters respectively while the proposed network contains only 75k parameters.

Overall, the proposed method is the third best performing algorithm in the challenging dataset of UBIRIS v2 while its complexity is at least an order of magnitude less than the two methods that outperforms it, making the proposed method more suited for deployment in embedded applications.

5.2.3.2. Comparison on CASIA Thousand and UBIRIS v2. In this section the proposed method is compared with other known segmentation methods over three important metrics: sensitivity, precision and F1-score. Sensitivity measures the model's ability to rule out non-iris pixels, while precision measures the ability of the model to detect true iris pixels. F1-score is the harmonic average of these two metrics. The segmentation methods include CAHT, GST, IFFP, OSIRIS, and WAHET. The comparisons are made over the CASIA Thousand and UBIRIS v2 original datasets. The numerical results are initially presented at (Hofbauer et al., 2014). The metrics for each presented algorithm are calculated comparing the algorithms results with the ground truth. The comparisons are illustrated in Figs. 24–25. Comparing all methods on the high-quality CASIA Thousand dataset, the proposed method achieves the best performance on the F1-score and precision metrics, and second best for the sensitivity metric. Furthermore, on UBIRIS v2 where the samples are of low quality the proposed method gives higher results in all metrics compared to the other approaches. Although the proposed method is designed for segmenting off-axis iris samples, these results show that it performs well on frontal iris samples of high and low quality.

6. Conclusion

In this paper advanced data augmentation techniques are proposed to simulate off-axis iris samples as represented when captured by user-facing cameras on wearable AR/VR headsets, which enables us to propose a low-complexity neural network architecture, designed for deployment on embedded devices, targeting the segmentation of off-axis iris samples. The current network represents a proof of concept which will be integrated into hardware in future works. The quality of segmentation achieved by

Table 7

Comparison between the proposed method and the SPDNN on the original samples from several datasets. Green colour shows a better performance. Yellow shows a marginal difference in the performance and Orange a noteworthy difference in performance. A higher value for μ and lower for σ is desired.

| Metrics		Bath800		CASIA Thousand		UBIRIS v2	
		Proposed Method	*SPDNN*	*Proposed Method*	*SPDNN*	*Proposed Method*	*SPDNN*
Accuracy	μ	99.13%	98.55%	99.50%	99.71%	98.92%	99.30%
	σ	0.52%	1.43%	0.36%	0.33%	0.67%	0.54%
Sensitivity	μ	94.90%	96.03%	94.67%	97.96%	88.38%	93.98%
	σ	6%	4.76%	4.33%	2.95%	9.29%	9.45%
Specificity	μ	99.56%	99.10%	99.86%	99.82%	99.71%	99.62%
	σ	0.47%	1.07%	0.16%	0.20%	0.39%	0.48%
Precision	μ	95.67%	96.05%	97.39%	97.13%	96.33%	94.88%
	σ	6.33%	4.46%	2.83%	3.10%	7.22%	5.40%
NPV	μ	99.49%	99.05%	99.63%	99.87%	99.10%	99.60%
	σ	0.45%	1.49%	0.34%	0.28%	0.60%	0.30%
F1-Score	μ	95.17%	95.93%	95.94%	97.50%	91.46%	93.90%
	σ	5.43%	3.88%	2.89%	2.51%	9.63%	9.70%
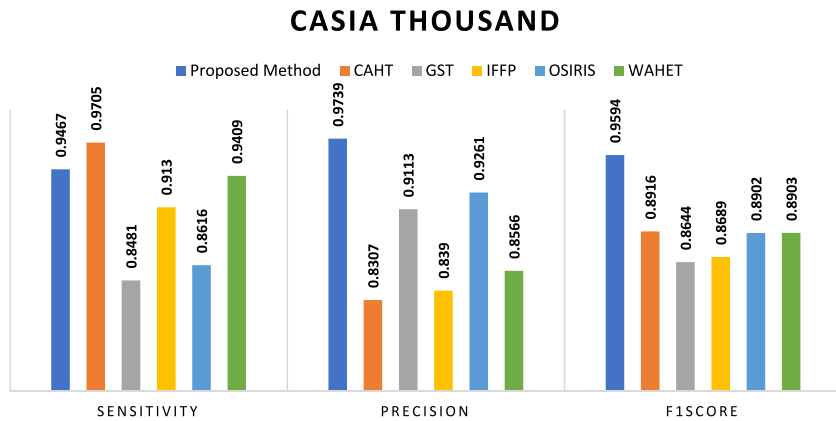
Fig. 24. Sensitivity, Precision, F1-score on the original samples of CASIA Thousand for the proposed method vs. five other methods.
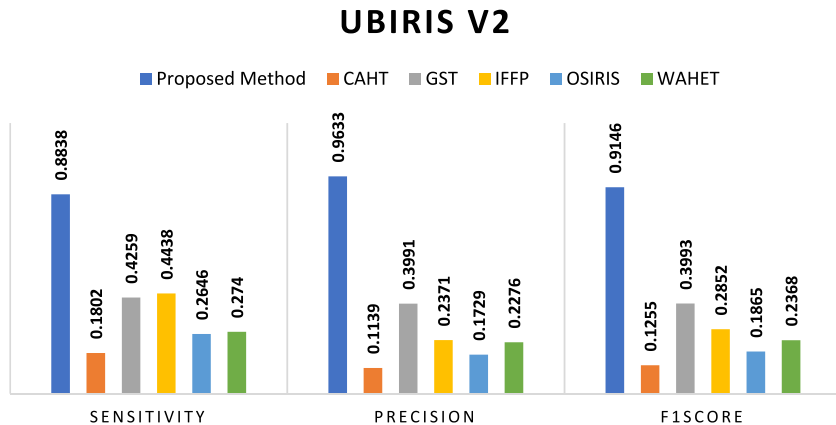

Fig. 25. Sensitivity, Precision, F1-score on the original samples of UBIRIS v2 for the proposed method vs. five other methods.

this network is evaluated and compared with state-of-the-art methods both for off-axis and frontal iris regions.

The proposed network's complexity is at least an order of magnitude less than other CNNs specifically designed for the iris segmentation task. Also, it has the best performance on segmenting the augmented off-axis iris samples. Further, the segmentation performance of this network on frontal iris samples from several public datasets, is comparable with the SPDNN network proposed by Bazrafkan et al. (2018) a state-of-the-art iris segmentation method. This performance is achieved even though the proposed network is of significantly smaller size and complexity and is trained for the task of segmenting off-axis iris samples. Due to its lightweight design and high performance in segmenting both off-axis and frontal iris samples and handling a range of input image qualities, the proposed network is well suited for general deployment on AR/VR devices.

Future work will focus on refinements in the network design and training/augmentation methodologies to improve performance on specific AR/VR headsets. As can be noted from the introduction, different devices will have user-facing cameras in a more limited set of locations and image acquisitions will be at varying NIR/wavelengths. In addition, the imaging pipeline on each camera module can have subtle effects on image quality.

Some practical examples of further research topics include developing an optimized CNN design based on SPDNN methods with a similar, or perhaps even smaller number of parameters that can achieve similar segmentation accuracy to our network. An additional further work includes the study of disease affected irises and the design of a CNN segmentation technique that is able to handle such iris images. Another future research direction is to build some device-specific datasets with iris images captured by the user-facing camera on several state-of-the-art AR/VR headsets. This will enable evaluation of the proposed segmentation method on practical off-axis iris samples. (At present it is not possible to gain low-level access to the imaging systems on the available devices to capture continuous image streams, but we have opened some discussions with device manufacturers and such access will hopefully be available in the near future as these devices continue to enter mainstream adoption.)

It is also expected to extend this work to apply these improved segmentation techniques to a number of full iris recognition pipelines to evaluate its effects on the reliability and robustness of near-view, off-axis iris recognition. The main challenge here is that the only off-axis recognition pipeline that we are aware of is proprietary. Again, we expect other algorithms will appear in the near future and hopefully some of these will be open-source or provide at least API-level access to system developers.

Acknowledgements

References

Abhyankar, A., Hornak, L., & Schuckers, S. (2005). Off-angle iris recognition using bi-orthogonal wavelet network system. In *Null* (pp. 239–244).

Abhyankar, A., & Schuckers, S. (2006). Active shape models for effective iris segmentation. In *Biometric technology for human identification III, Vol. 6202* (p. 62020H).

Ackerman, E. (2013). Google gets in your face: Google glass offers a slightly augmented version of reality. *IEEE Spectrum*, *50*(1), 26–29. http://dx.doi.org/10.1109/MSPEC.2013.6395302.

Arsalan, M., Hong, H. G., Naqvi, R. A., Lee, M. B., Kim, M. C., Kim, D. S., et al. (2017). Deep learning-based iris segmentation for iris recognition in visible light environment. *Symmetry*, *9*(11), 263.

Arsalan, M., Naqvi, R. A., Kim, D. S., Nguyen, P. H., Owais, M., & Park, K. R. (2018). Irisdensenet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light Camera sensors. *Sensors*, *18*(5), 1501.

Bakir, A., Chesler, G., & Torriente, M. de la (2016). *Using touch ID for local authentication. Internet of things with swift for IOS*.

Bazrafkan, S., Thavalengal, S., & Corcoran, P. (2018). An end to end deep neural network for iris segmentation in unconstrained scenarios. *Neural Networks*, *106*, 79–95. http://dx.doi.org/10.1016/J.NEUNET.2018.06.011.

Bhorkar, G. (2017). A survey of augmented reality navigation. *Foundations and Trends® in Human–Computer Interaction*, *8*(2–3), 73–272. http://dx.doi.org/10.1561/1100000049.

Bowyer, K. W., Hollingsworth, K., & Flynn, P. J. (2008). Image understanding for iris biometrics: A survey. *Computer Vision and Image Understanding*, *110*(2), 281–307.

Bowyer, K. W., Hollingsworth, K. P., & Flynn, P. J. (2013). A survey of iris biometrics research: 2008–2010. In *Handbook of iris recognition, Vol. 1* (pp. 15–54). Springer.

Broussard, R. P., & Ives, R. W. (2009). Using artificial neural networks and feature saliency to identify iris measurements that contain the most discriminatory information for iris segmentation. In *Computational intelligence in biometrics: Theory, algorithms, and applications, 2009. CIB 2009. IEEE workshop on* (pp. 46–51).

CASIA Iris Image Database (2019). (n.d.) Retrieved from http://biometrics.idealtest.org/.

Cave, A. (2015). Why Google Glass Flopped. Retrieved from http://www.forbes.com/sites/andrewcave/2015/01/20/a-failure-of-leadership-or-design-why-google-glass-flopped/#24c650d3556a.

Chan, P., Halevi, T., & Memon, N. (2015). *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics, Glass OTP: Secure and convenient user authentication on google glass*. http://dx.doi.org/10.1007/978-3-662-48051-9_22.

Chauhan, J., Asghar, H. J., Kâafar, M. A., & Mahanti, A. (2016). Gesture-based continuous authentication for wearable devices: the google glass Case. In *14th international conference on applied cryptography and network security*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. ArXiv Preprint ArXiv:1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. ArXiv Preprint ArXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 801–818).

Cherapau, I., Muslukhov, I., Arachchilage, N. A. G., & Beznosov, K. (2015). On the impact of touch ID on iphone passcodes. In *Symposium on usable privacy and security (SOUPS)* (pp. 257–276).

Ching, K. W., & Singh, M. M. (2016). Wearable technology devices security and privacy vulnerability analysis. *International Journal of Network Security & Its Applications*, http://dx.doi.org/10.5121/ijnsa.2016.8302.

Cognard, Timothée E., Goncharov, Alexander, Devaney, Nicholas, Dainty, Chris, & Corcoran, Peter (2018). undefined. (n.d.) A Review of Resolution Losses for AR/VR Foveated Imaging Applications. Ieeexplore.Ieee.Org.

Corcoran, P. M. (2013). Biometrics and consumer electronics: A brave new world or the road to dystopia?. *Consumer Electronics Magazine IEEE*, *2*(2), 22–33.

Corcoran, Peter (2016). The battle for privacy in your pocket [notes from the editor]. *IEEE Consumer Electronics Magazine*, *5*(3), 3–36. http://dx.doi.org/10.1109/MCE.2016.2558218.

Corcoran, P. M. (2017). A privacy framework for the internet of thing. In *2016 IEEE 3rd world forum on internet of things, WF-IoT 2016*. http://dx.doi.org/10.1109/WF-IoT.2016.7845505.

Corcoran, Peter, Bigioi, P., & Thavalengal, S. (2015). Feasibility and design considerations for an iris acquisition system for smartphones. In *IEEE international conference on consumer electronics - Berlin, ICCE-Berlin (Vol. 2015-Febru* (pp. 164–167). IEEE, http://dx.doi.org/10.1109/ICCE-Berlin.2014.7034328.

Corcoran, P., & Costache, C. (2016). Biometric technology and smartphones: A consideration of the practicalities of a broad adoption of biometrics and the likely impacts. *IEEE Consumer Electronics Magazine*, *5*(2), 70–78. http://dx.doi.org/10.1109/MCE.2016.2521937.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. ArXiv Preprint ArXiv:1805.09501.

Darwaish, S. F., Moradian, E., Rahmani, T., & Knauer, M. (2014). Biometric identification on android smartphones. In *Procedia computer science, Vol. 35* (pp. 832–841). http://dx.doi.org/10.1016/j.procs.2014.08.250.

Daugman, J. (2007). New methods in iris recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 37(5), 1167–1175.

Daugman, J. (2009). How iris recognition works. In *The essential guide to image processing* (pp. 715–739). Elsevier.

De Luca, A., Hang, A., von Zezschwitz, E., & Hussmann, H. (2015). I feel like i'm taking selfies all day!. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems - CHI '15*. http://dx.doi.org/10.1145/2702123.2702141.

Dorairaj, V., Schmid, N. A., & Fahmy, G. (2005). Performance evaluation of non-ideal iris based recognition system implementing global ICA encoding. In *Image processing, 2005. ICIP 2005. IEEE international conference on, Vol. 3*. III–285.

Elise, B. (2014). Google Glass a Game-changing Application in the Realm of Cultural Tourism. Business Wire. Retrieved from http://search.proquest.com/docview/1635054301?accountid=14744%5Cnhttp://fama.us.es/search*spi/i?SEARCH=%5Cnhttp://pibserver.us.es/gtb/usuario_acceso.php?centro=USEG&centro=%24USEG&d=1.

Erbilek, M., Da Costa-Abreu, M. C., & Fairhurst, M. (2012). *Optimal configuration strategies for iris recognition processing*.

Fox, B., & Felkey, B. (2013). Potential uses of google glass in the pharmacy. *Hospital Pharmacy*, 48(9), 783–784. http://dx.doi.org/10.1310/hpj4809-783.

Gangwar, A., Joshi, A., Singh, A., Alonso-Fernandez, F., & Bigun, J. (2016). IrisSeg: A fast and robust iris segmentation framework for non-ideal iris images. In *2016 international conference on biometrics (ICB)* (pp. 1–8). IEEE.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. 2017. A review on deep learning techniques applied to semantic segmentation. ArXiv Preprint ArXiv:1704.06857.

Goode, A. (2014). Bring your own finger – how mobile is bringing biometrics to consumers. *Biometric Technology Today*, 2014(5), 5–9. http://dx.doi.org/10.1016/S0969-4765(14)70088-8.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hammal, Z., Massot, C., Bedoya, G., & Caplier, A. (2005). In S. Singh, M. Singh, C. Apte, & P. Perner (Eds.), *Eyes segmentation applied to gaze direction and vigilance estimation BT - pattern recognition and image analysis* (pp. 236–246). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hayes, A. (2016). My journey into glass: Talking about google glass with stakeholders in the glass explorer program. *IEEE Consumer Electronics Magazine*, 5(1), 102–105. http://dx.doi.org/10.1109/MCE.2015.2484819.

He, X., & Shi, P. (2006). A novel iris segmentation method for hand-held capture device. In *International conference on biometrics* (pp. 479–485). Springer.

Hofbauer, Heinz, Alonso-Fernandez, F., Bigun, J., & Uhl, A. (2016). Experimental analysis regarding the influence of iris segmentation on the recognition rate. *IET Biometrics*, 5(3), 200–211.

Hofbauer, H., Alonso-Fernandez, F., Wild, P., Bigun, J., & Uhl, A. (2014). A ground truth for iris segmentation. In *2014 22nd international conference on pattern recognition* (pp. 527–532). http://dx.doi.org/10.1109/ICPR.2014.101.

Huang, Y.-P., Luo, S.-W., & Chen, E.-Y. (2002). An efficient iris recognition system. In *Machine learning and cybernetics, 2002. Proceedings. 2002 international conference on, Vol. 1* (pp. 450–454). IEEE.

Jalilian, E., Uhl, A., & Kwitt, R. (2017). *Domain adaptation for CNN based Iris segmentation. BIOSIG 2017*.

Jan, F. (2017). Segmentation and localization schemes for non-ideal iris biometric systems. *Signal Processing*, 133, 192–212.

Jiang, Z., Yuan, Y., & Wang, Q. (2018). Contour-aware network for semantic segmentation via adaptive depth. *Neurocomputing*, 284, 27–35.

Jillela, R., & Ross, A. A. (2013). Methods for iris segmentation. In *Handbook of Iris recognition* (pp. 239–279). Springer.

Khan, T. M., Khan, M. A., Malik, S. A., Khan, S. A., Bashir, T., & Dar, A. H. (2011). Automatic localization of pupil using eccentricity and iris using gradient based method. *Optics and Lasers in Engineering*, 49(2), 177–187.

Koh, J., Govindaraju, V., & Chaudhary, V. (2010). A robust iris localization method using an active contour model and hough transform. In *Pattern recognition (ICPR), 2010 20th international conference on* (pp. 2852–2856).

Kress, B., Saeedi, E., & Brac-de-la Perriere, V. (2014). The segmentation of the HMD market: optics for smart glasses, smart eyewear, AR and VR headsets. In *Photonics applications for aviation, aerospace, commercial, and harsh environments V, Vol. 9202* (p. 92020D). http://dx.doi.org/10.1117/12.2064351.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lakra, A., Tripathi, P., Keshari, R., Vatsa, M., & Singh, R. (2018). Segdensenet: Iris segmentation for pre-and-post Cataract surgery. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 3150–3155). IEEE.

Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*.

Lemley, J., Bazrafkan, S., & Corcoran, P. (2017). Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, 5, 5858–5869. http://dx.doi.org/10.1109/ACCESS.2017.2696121.

Li, X. (2006). Modeling intra-class variation for nonideal iris recognition. In *International conference on biometrics* (pp. 419–427).

Lili, P., & Mei, X. (2005). The algorithm of iris image preprocessing. In *Automatic identification advanced technologies, 2005. Fourth IEEE workshop on* (pp. 134–138). IEEE.

Linao, M. (2016). *The present and future of VR/AR: Applications in education, gaming, commerce, and industry*.

Liu, N., Li, H., Zhang, M., Liu, J., Sun, Z., & Tan, T. (2016). Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In *Biometrics (ICB), 2016 International conference on* (pp. 1–8).

Liu, Y., Yuan, S., Zhu, X., & Cui, Q. (2003). A practical iris acquisition system and a fast edges locating algorithm in iris recognition. In *IEEE instrumentation and measurement technology conference proceedings, Vol. 1* (pp. 166–169).

Mann, S. (2001). Fundamental issues in mediated reality, wearcomp, and camera-based augmented reality. In *Fundamentals of wearable computers and augmented reality* (pp. 295–328). Lawrence Erlbaum Associates, Inc., http://dx.doi.org/10.1061/(ASCE)0733-947X(2005)131:3(169).

Mann, S. (2004). Continuous lifelong capture of personal experience with EyeTap. In *Proceedings of the 1st ACM workshop on continuous archival and retrieval of personal experiences - CARPE'04* (pp. 1–21). http://dx.doi.org/10.1145/1026653.1026654.

Mann, S. (2013). Steve mann: My augmented life. *IEEE Spectrum*, 1–6.

Mann, S., & Fung, J. (2002). Eyetap devices for augmented, deliberately diminished, or otherwise altered visual perception of rigid planar patches of real-world scenes. *Presence: Teleoperators & Virtual Environments*, 11(2), 158–175. http://dx.doi.org/10.1162/1054746021470603.

MIRLIN (2019). (n.d.) Retrieved from https://www.fotonation.com/products/biometrics/iris-recognition/.

Muensterer, O. J., Lacher, M., Zoeller, C., Bronstein, M., & Kübler, J. (2014). Google glass in pediatric surgery: An exploratory study. *International Journal of Surgery*, 12(4), 281–289. http://dx.doi.org/10.1016/j.ijsu.2014.02.003.

Othman, N., Dorizzi, B., & Garcia-Salicetti, S. (2016). OSIRIS: An open source iris recognition software. *Pattern Recognition Letters*, 82, 124–131.

Peng, G., Zhou, G., Nguyen, D. T., Qi, X., Yang, Q., & Wang, S. (2017). Continuous authentication with touch behavioral biometrics and voice on wearable glasses. *IEEE Transactions on Human-Machine Systems*, http://dx.doi.org/10.1109/THMS.2016.2623562.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. ArXiv Preprint ArXiv:1712.04621.

Prabhakar, S., Pankanti, S., & Jain, A. K. (2003). Biometric recognition: security and privacy concerns. *IEEE Security & Privacy*, 1(2), 33–42. http://dx.doi.org/10.1109/MSECP.2003.1193209.

Proenca, H. (2010). Iris recognition: On the segmentation of degraded images acquired in the visible wavelength. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 1502–1516.

Proenca, H., Filipe, S., Santos, R., Oliveira, J., & Alexandre, L. A. (2010). The ubiris.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 1529–1535.

Proença, H., & Alexandre, L. A. (2010). Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage. *Image and Vision Computing*, 28(1), 202–206.

Quinn, G. W., Grother, P. J., Ngan, M. L., & Matey, J. R. (2013). IREX IV: part 1, evaluation of iris identification algorithms.

Radman, A., Zainal, N., & Suandi, S. A. (2017). Automated segmentation of iris images acquired in an unconstrained environment using HOG-svm and growcut. *Digital Signal Processing* , 64, 60–70.

Rakshit, S. (2007). *Novel methods for accurate human Iris recognition*. University of Bath.

Ring, T. (2015). Spoofing: are the hackers beating biometrics?. *Biometric Technology Today*, 2015(7), 5–9. http://dx.doi.org/10.1016/S0969-4765(15)30119-3.

Rompapas, D. C., Rovira, A., Ikeda, S., Plopski, A., Taketomi, T., Sandor, C., et al. (2017). Eyear: Refocusable augmented reality content through eye measurements. In *Adjunct proceedings of the 2016 IEEE international symposium on mixed and augmented reality, ISMAR-Adjunct 2016*. http://dx.doi.org/10.1109/ISMAR-Adjunct.2016.0108.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3), 279–283.

Samangouei, P., Patel, V. M., & Chellappa, R. (2017). Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, http://dx.doi.org/10.1016/j.imavis.2016.05.004.

Schlüter, J., & Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR* (pp. 121–126).

Schreinemacher, M. H., Graafland, M., & Schijven, M. P. (2014). Google glass in surgery. *Surgical Innovation*, http://dx.doi.org/10.1177/1553350614546006.

Shah, S., & Ross, A. (2009). Iris segmentation using geodesic active contours. *IEEE Transactions on Information Forensics and Security*, *4*(4), 824–836.

Shejin, Thavalengal, & Corcoran, P. (2016). User authentication on smartphones: Focusing on iris biometrics. *IEEE Consumer Electronics Magazine*, *5*(2), 87–93. http://dx.doi.org/10.1109/MCE.2016.2522018.

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). Research on data augmentation for image classification based on convolution neural networks. In *2017 Chinese automation congress (CAC)* (pp. 4165–4170). IEEE.

Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., et al. (1997). Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments*, *6*(4), 386–398. http://dx.doi.org/10.1162/pres.1997.6.4.386.

Tan, T., He, Z., & Sun, Z. (2010). Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition. *Image and Vision Computing*, *28*(2), 223–230.

Tan, C.-W., & Kumar, A. (2012). Unified framework for automated iris segmentation using distantly acquired face images. *IEEE Transactions on Image Processing*, *21*(9), 4068–4079.

Tan, C.-W., & Kumar, A. (2013). Towards online iris and periocular recognition under relaxed imaging constraints. *IEEE Transactions on Image Processing*, *22*(10), 3751–3765.

Tang, F., Aimone, C., Fung, J., Marjan, A., & Mann, S. (2002). Seeing eye to eye: A shared mediated reality using eyetap devices and the videoorbits gyroscopic head tracker. In *Proceedings - international symposium on mixed and augmented reality, ISMAR 2002* (pp. 267–268). http://dx.doi.org/10.1109/ISMAR.2002.1115106.

Taylor, L., & Nitschke, G. (2017). Improving deep learning using generic data augmentation. ArXiv Preprint ArXiv:1708.06020.

Thavalengal, Shejin, Andorko, I., Drimbarean, A., Bigioi, P., & Corcoran, P. (2015). Proof-of-concept and evaluation of a dual function visible/NIR camera for iris authentication in smartphones. *IEEE Transactions on Consumer Electronics*, *61*(2), 137–143. http://dx.doi.org/10.1109/TCE.2015.7150566.

Thavalengal, S., Bigioi, P., & Corcoran, P. (2015a). Evaluation of combined visible/NIR camera for iris authentication on smartphones. In *2015 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 42–49). http://dx.doi.org/10.1109/CVPRW.2015.7301318.

Thavalengal, S., Bigioi, P., & Corcoran, P. (2015b). Iris authentication in handheld devices - considerations for constraint-free acquisition. *IEEE Transactions on Consumer Electronics*, *61*(2), 245–253. http://dx.doi.org/10.1109/TCE.2015.7150600.

Thavalengal, Shejin, Bigioi, P., & Corcoran, P. (2016). Efficient segmentation for multi-frame iris acquisition on smartphones. In *2016 IEEE international conference on consumer electronics (ICCE) (2016 ICCE)* (pp. 202–203). Las Vegas, USA.

Timekeeper (2017). The promise of augmented reality. *The Economist*.

Tipton, Stephen J., White II, Daniel J., Sershon, Christopher, & Choi, Young B. (2014). Ios security and privacy: Authentication methods, permissions, and potential pitfalls with touch id. *International Journal of Computer and Information Technology*, *3*(3), 482–489.

Varkarakis, V., Bazrafkan, S., & Corcoran, P. (2018). A deep learning approach to segmentation of distorted iris regions in head-mounted displays. In *2018 IEEE games, entertainment, media conference (GEM)* (pp. 1–9). IEEE.

Vazquez-Fernandez, E., & Gonzalez-Jimenez, D. (2016). Face recognition for authentication on mobile devices. *Image and Vision Computing*, http://dx.doi.org/10.1016/j.imavis.2016.03.018.

Wang, Q., Gao, J., & Yuan, Y. (2018). Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. *IEEE Transactions on Intelligent Transportation Systems*, *19*(1), 230–241.

WaveLab (2019). (n.d.) No Title..

Wildes, R. P. (1997). Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, *85*(9), 1348–1363.

Yadav, D. K., Ionascu, B., Ongole, S. V. K., Roy, A., & Memon, N. (2015). *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics, Design and analysis of shoulder surfing resistant pin based authentication mechanisms on google glass*. http://dx.doi.org/10.1007/978-3-662-48051-9_21.

Zhao, Z., & Ajay, K. (2015). An accurate iris segmentation framework under relaxed imaging constraints using total variation model. In *Proceedings of the IEEE international conference on computer vision* (pp. 3828–3836).

Appendix B

A Deep Learning approach to Segmentation of Distorted Iris regions in Head-Mounted Displays

A Deep Learning Approach to Segmentation of Distorted Iris Regions in Head-Mounted Displays

3 authors:

Viktor Varkarakis
National University of Ireland, Galway
15 PUBLICATIONS   **43** CITATIONS

SEE PROFILE

Shabab Bazrafkan
University of Antwerp
44 PUBLICATIONS   **624** CITATIONS

SEE PROFILE

Peter Corcoran
National University of Ireland, Galway
608 PUBLICATIONS   **3,998** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Deep Iris Segmentation View project

Project   Off-axis Iris Segmentation for AR/VR with Deep Neural Networks View project

A Deep Learning approach to Segmentation of Distorted Iris regions in Head-Mounted Displays

Viktor Varkarakis
Collage of Engineering and Informatics National University of Ireland Galway
Galway, Ireland
Email: v.varkarakis1@nuigalway.ie

Shabab Bazrafkan
Collage of Engineering and Informatics National University of Ireland Galway
Galway, Ireland
Email: s.bazrafkan1@nuigalway.ie

Peter Corcoran
Collage of Engineering and Informatics National University of Ireland Galway
Galway, Ireland
Email: peter.corcoran@nuigalway.ie

Abstract—In this paper, we consider the next generation of wearable AR/VR display glasses and the challenges of personal authentication on such devices. The use of iris authentication as a mean of creating a seamless biometric link between the user and his personal data offers a viable approach, but due to the likely location of user-facing cameras there are some challenges in achieving an accurate segmentation of the iris. In this paper, a deep neural network was trained to accurately segment distorted iris regions. An appropriate augmentation method is presented to generate the distorted iris dataset used for training from publicly available frontal iris datasets. The proposed method shows promising results in segmenting off-axis iris images in unconstrained conditions.

I. INTRODUCTION

Virtual reality display technology has only recently appeared as a consumer electronics product in the form of new VR headsets, but in industry sectors these headsets have been evolving for more than a decade [1]. But now that the technology is ready for mass market deployment it will not take long for the next generation of headset technology to evolve, providing even more sophisticated display and interface capabilities and acting as a gateway into new virtual and augmented application frameworks [2]–[5].

Now the current evolution in AR display systems does not imply that this is a 'new' technology. Researchers have in fact been working with Augmented Displays for more than 20 years [6]–[12]. The Microsoft HoloLens [3], [13]–[15] is a good example of an AR headset that is similar to today's VR headsets – perhaps a little oversized for day-to-day consumer use. The most recent mass market experiment with a wearable, augmented/mediated-reality display, that could be worn on a day-to-day basis, was Google Glass[11],[16], [17]. Glass, as it became known, was considered to be a game changing technology for a few years across a wide range of industry sectors [18]–[21]. Ultimately, however, the product was withdrawn [22] and since that point consumers have had to wait for the next mass-market AR display technology.

Currently there are several companies that are working on the next generation of wearable technology to follow in the footsteps of Google Glass and Steve Mann's EyeTap. There are numerous challenges both in the display technology itself, but also in the ergonomics and the user-interface aspects of the device. Nevertheless, we expect that recent advances in motion-sensing technologies, eye-tracking and affective interpretation models will improve the usability of the next generation of these devices.

Another key challenge which applies to all variants of AR and VR headsets is that of user authentication. Where there is no physical keyboard or equivalent there are challenges to authenticate a user and ensure the privacy of user profiles and data that is collected/generated by a wearable AR consumer device. It is interesting that in the recently released movie "Ready Player One" the players who enter the virtual world of "The Oasis" are portrayed as being linked auto-magically with their virtual world profiles, but no explanation of this magic "authentication" achieved, is provided. Fortunately, we can answer this question for you in this short paper by looking at how AR glasses are evolving today and demonstrating how current biometric authentication technology can be easily re-purposed to work in a seamless manner on tomorrow's wearable AR/VR glasses and display devices.

A good starting point is found in [23] where the author considers how biometric technology is becoming the natural authentication mechanism for personal consumer devices and the broad range of services and capabilities they bring to our daily lives. In follow-on publications the use of iris recognition on consumer devices is explored [24]–[27] and the importance of accurate iris segmentation, particularly in consumer imaging devices, is identified as a key challenge [28], [29]. In the iris authentication workflow, failed segmentations represent the single largest source of error [30]–[32].

In addition to its role in improving the performance of an iris-based authentication system, the accurate segmentation of iris regions, can be used successfully for eye-gaze estimation [33]. Eye-gaze is a key element of various user-interface modalities for wearable AR & VR displays.

As shown in US design patent, D795952S1 [34] "Fig1", a possible location of the camera used for iris authentication or eye-gaze direction, is below the eye. As opposed to when the camera is frontal, resulting in having circular iris images, in positioning the camera, same way or similar as described above, off-axis iris images will be obtained. At the moment

there is not a publicly available commercial segmentation method for off-axis iris images as this is quite a new problem arising from the recent proliferation of AR/VR technology into consumer devices.
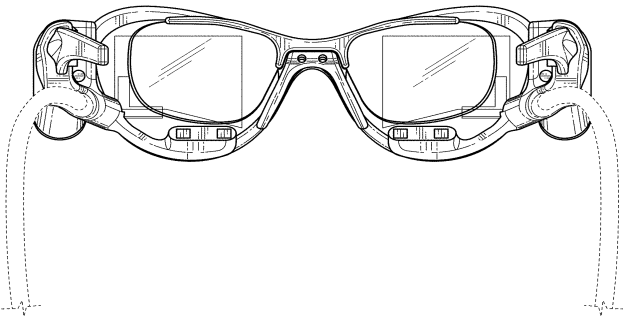

Fig. 1. Virtual Reality Glasses - Patent Number: USD795952S1 [34]

The focus of this work is on implementing a deep learning technique to segment distorted iris images. The main contribution is a data augmentation technique that simulates iris images from an AR/VR head mounted display with off-axis cameras. On a selection of the distorted iris images a second augmentation process is applied, adding contrast, blurring or shadows to the images mimicking the variability of images quality achieved in real-life environment.

In the next section the proposed method is explained in more detail including the data preparation and augmentation techniques that are implemented. The design of the deep learning network is also presented along with a description of the underlying training algorithm. The results of the proposed method are given in the final section.

II. PROPOSED METHOD

A. Database Preparation

In order to accurately train a deep neural network, a large number of labeled training samples are required. These samples should correctly characterize the imaging problem so that it can be solved, enabling the deep learning process to train an accurate model. For this reason, an augmentation process was implemented and is explained in section: B. Data Augmentation. The CASIA Thousand [35] and Bath 800 [36] databases were used as a starting point for the task of generating a suitably augmented training dataset. The CASIA Thousand has approximately 20.000 images and Bath 800 over 30.000 iris images. These databases are not labeled with the segmentation ground truth, but they do contain high quality iris images (captured in constrained conditions). There are high resolution, high contrast samples, with low noise and shadowing. These high-quality images can be accurately segmented with standard industrial segmentation algorithms applied to segment the original images. In this work, these segmentations are considered as the effective ground truth for the iris images.

B. Data Augmentation

To proceed to the training of a deep neural network for AR/VR iris segmentation task, an augmentation of the database

is required as there is not a publicly available database of iris images acquired from an AR/VR set-up. Therefore, the first objective of the augmentation process is to simulate the representation of respective iris images. This representation, consist of distorted iris images, some with secondary low contrast, blurring and shadowing of the distorted iris images. In this work, all the samples are resized to 120x160 using bicubic interpolation. To generate the initial distorted iris images the samples were warped using two transformations.

1. The images are warped by applying a spatial stretching/contacting of the iris images in different parts. For example, an image will be stretched in its left side and contracted in its right side as shown in "Fig. 2". The warping is applied in a linear manner with random parameters. The parameters determine the amount of stretching/contracting of the image in each direction. The void spaces are filled using linear interpolation. This transformation is applied in both horizontal and vertical direction. After that the images are resized to the original resolution (120x160). Applying warping results in iris images with non-circular iris and pupil structures, as shown in "Fig.3" which is a usual case in iris images acquired from an AR/VR headset. This transformation is applied with 50% probability to the images.

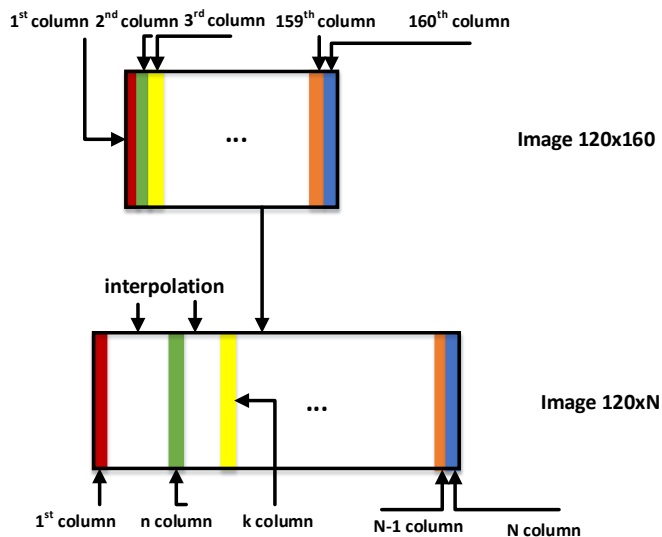Fig. 2. Stretching/Contracting of the image.

Original Image Ground Truth Warped Image Warped Ground Truth
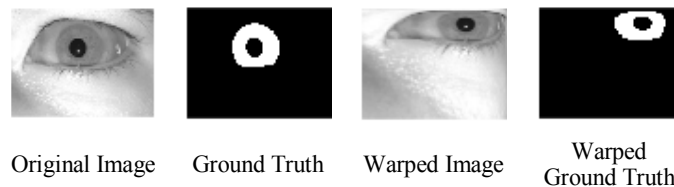
Fig. 3. Example of a warped image.

2. At the second transformation, the images are tilted in two directions (up-left, up-right) with the same probability.

This is accomplished by applying a projective transformation to the images. This transformation is mapping the top vertices of the image to a new pair of point as illustrated in "Fig. 4". In "Fig. 4" the values of a, b, c and d are randomly generated so that the image is tilted in the desired direction. The values from "Fig.4", of a and c are in U (0.9,1), b is in U (0.15,0.45) and d in U (0.55,0.85) where U is the uniform distribution. The images were tilted in these directions, due to the fact that, as shown in "Fig. 1" and described in the introduction, a possible location of the camera used for capturing the iris images will be below the eye. The second transformation is applied to the images that the first transformation was not applied, along with the images that were distorted in the previous step with 50% probability. In "Fig.5" an image is shown where the above transformation was applied.
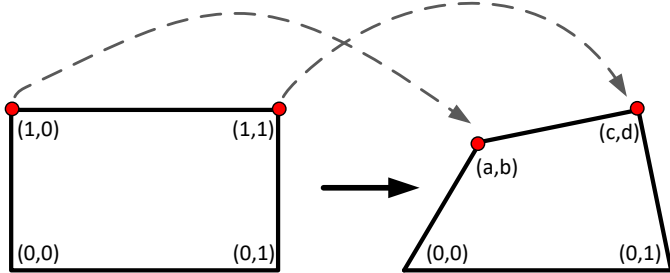

Fig. 4. How the projective transformation is applied to an image.

Original Image   Ground Truth   Tilted Image   Tilted Ground Truth

Fig. 5. Example of tilted image.

The second objective of the augmentation, is to ensure that the samples used to train the network represent real-life scenarios. The distribution of the input data plays an extremely important role in what the network learns and how will behave during the testing stage but also in unconstrained situations [37]. The databases that are used, CASIA Thousand and Bath 800, consist of high quality iris images. Based on the precise observations that have been done [26], [38] the main differences between high quality constrained images and wild ones, are linked to contrast, blurring and shadows in the image. In order to simulate real-life captured iris images, the contrast inside and outside the iris region is changed separately using histogram mapping. Motion blurring is applied, as well as, adding shadows to the images by multiplying them with a shadow function. Contrast and blurring are applied to all the images and with 50% probability shadowing is applied to the samples. The functions used in order to apply contrast, shadowing and motion blurring are explained in [29]. The augmentation process is shown in "Fig. 6". The distortion-contrast database and the distortion database, as illustrated in "Fig. 6", are created twice and the contrast database once.

After the augmentation process, with the addition of the original images the total number of samples generated is over

300.000. Regarding the ground truth of the augmented data, when a distortion method is applied to an image, an identical distortion is also applied to its ground truth segmentation. In "Fig.7" examples of augmented data are presented along with their corresponding augmented ground truth.
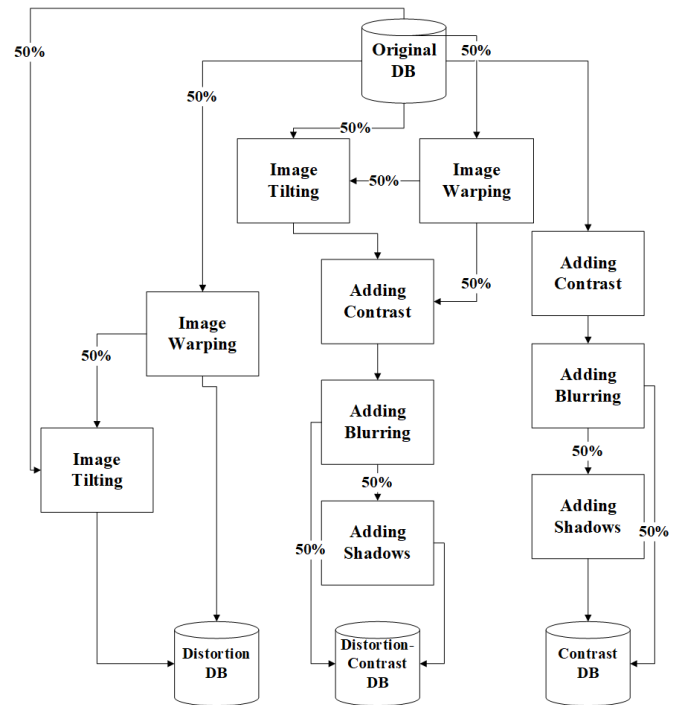
Fig. 6. Workflow of the augmentation process.
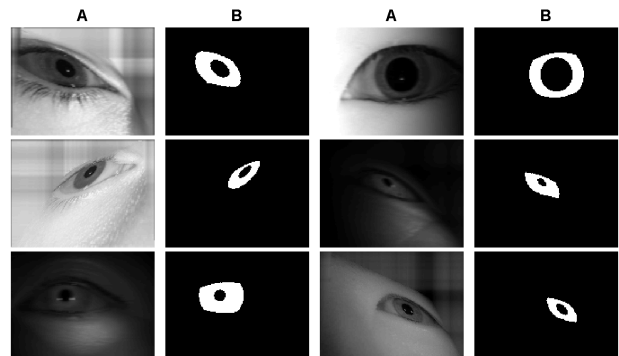
Fig. 7. A: Augmented samples, B: Augmented Ground Truth of Segmentation

C. Network Design

For the segmentation task a fully convolutional network inspired by [29] is used, consisting of 10 layers. The network starts with a 3x3 kernel mapping the input (1 channel) on the first convolutional hidden layer which consists of 32 channels using a rectified linear unit (ReLu) as an activation function. The kernel size remains the same throughout the convolutional hidden layers, as well as, the number of channels and their activation function. Finally, at the output layer (1 channel), the kernel size is 3x3, but in this layer, the sigmoid activation function is used. Pooling layers were not used as it was

observed that the accuracy of the network's output was decreasing.

D. Training

The training was carried out in TensorFlow library. The Mean Squared Error is used as the loss function. The Gradient Descent with Adaptive Moment Estimation (Adam) is used, with a learning rate of 1e-4, beta1 and beta2 equal to 0.9 and 0.999 respectively, in order to optimize the loss function. The samples are divided 70% for the training set, 20% for validation set and 10% for test set. The training is done on a desktop computer with Nvidia GTX 1080 GPU.

III. RESULTS AND DISCUSSION

In this work, the network is trained on the original images and the augmented databases (Bath 800 and CASIA Thousand). The output of the network is a grayscale segmentation map with values between 0 and 1. The binary map is obtained by using a thresholding technique, where the values bigger than the threshold are shifted to 1 and the others to 0. The threshold value 0.55 is used in this work. In "Fig.8 the output of the network is presented for several sample images. Several metrics have been used to evaluate the network. The metrics used are described thoroughly in [29].

The segmentation results for the test set on the two databases (Bath 800 and CASIA Thousand) are presented below in Table I and II. In Table I, higher performance is represented by higher values and in Table II higher performance is represented by lower values.

The proposed method shows promising results using the deep learning technique in segmenting off-axis iris images as represented by AR/VR set-ups, including also effects on the images from real-life environments.

This work is an initial proof-of-concept and more details regarding the augmentation process along with numerical analysis and comparisons with other segmentation algorithms will be presented at the conference.

TABLE I. SEGMENTATION RESULTS

Metrics	Proposed Method	
	BATH 800	CASIA Thousand
Accuracy	99.12%	99.34%
Sensitivity	92.75%	90.32%
Specificity	99.58%	99.78%
Precision	94.26%	94.78%
NPV	99.45%	99.52%
F1-Score	93.23%	92.07%
MCC	92.94%	92.02%
Informedness	92.34%	90.10%
Markedness	93.72%	94.30%

TABLE II. SEGMENTATION RESULTS

Metrics	Proposed Method	
	BATH 800	CASIA Thousand
FPR	0.41%	0.22%
FNR	7.24%	9.67%
FDR	5.73%	5.21%
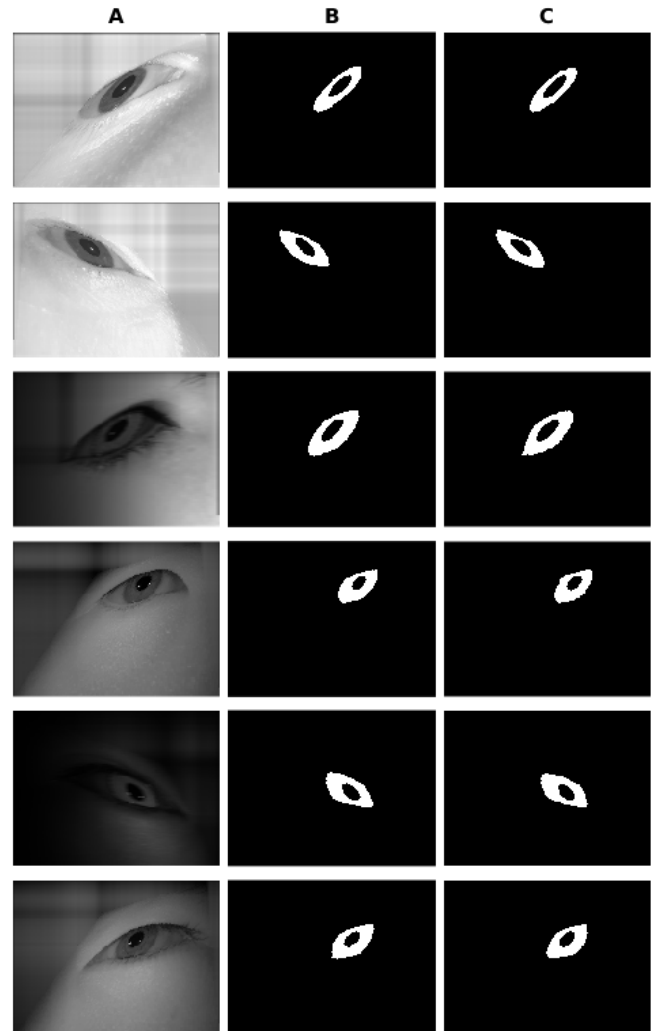
Fig. 8. A: Augmented samples, B: Segmentation Ground Truth, C: Output of Network - Segmentation map

REFERENCES

[1]   B. Kress, E. Saeedi, and V. Brac-de-la-Perriere, "The segmentation of the HMD market: optics for smart glasses, smart eyewear, AR and VR headsets," in *Photonics Applications for Aviation, Aerospace, Commercial, and Harsh Environments V*, 2014, vol. 9202, p. 92020D.

[2]   D. De Angeli and E. J. O'Neill, "Development of an Inexpensive Augmented Reality (AR) Headset," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*, 2015, pp. 971–976.

[3] R. Furlan, "The future of augmented reality: Hololens - Microsoft's AR headset shines despite rough edges [Resources-Tools and Toys]," *IEEE Spectr.*, vol. 53, no. 6, p. 21, 2016.

[4] M. Linao, "The Present And Future Of VR/AR: Applications In Education, Gaming, Commerce, And Industry," *CB Insights*, 2016.

[5] D. Wagner and D. Shmalstieg, "Making AR practical on Mobile Phones, Part 1," *IEEE Computer Graphics and Applications*, pp. 12–15, 2009.

[6] T. Starner *et al.*, "Augmented Reality through Wearable Computing," *Presence Teleoperators Virtual Environ.*, vol. 6, no. 4, pp. 386–398, 1997.

[7] S. Mann, "Fundamental issues in mediated reality, WearComp, and camera-based augmented reality," *Fundam. Wearable Comput. Augment. Reality, Lawrence Erlbaum Assoc. Inc*, pp. 295–328, 2001.

[8] S. Mann and J. Fung, "EyeTap Devices for Augmented, Deliberately Diminished, or Otherwise Altered Visual Perception of Rigid Planar Patches of Real-World Scenes.," *Presence Teleoperators Virtual Environ.*, vol. 11, no. 2, pp. 158–175, 2002.

[9] S. Mann, "Continuous lifelong capture of personal experience with EyeTap," *Proc. 1st ACM Work. Contin. Arch. Retr. Pers. Exp. - CARPE'04*, pp. 1–21, 2004.

[10] F. Tang, C. Aimone, J. Fung, A. Marjan, and S. Mann, "Seeing eye to eye: A shared mediated reality using EyeTap devices and the VideoOrbits gyroscopic head tracker," in *Proceedings - International Symposium on Mixed and Augmented Reality, ISMAR 2002*, 2002, pp. 267–268.

[11] S. Mann, "Steve Mann: My 'Augmented' Life," *IEEE Spectr.*, pp. 1–6, 2013.

[12] G. Bhorkar, "A Survey of Augmented Reality Navigation," *Found. Trends® Human–Computer Interact.*, vol. 8, no. 2–3, pp. 73–272, 2017.

[13] M. Fitzsimmons, "Hands on: Microsoft HoloLens review," *techradar*, 2016. [Online]. Available: http://www.techradar.com/reviews/wearables/microsoft-hololens-1281834/review.

[14] M. Garon, P.-O. Boulet, J.-P. Doironz, L. Beaulieu, and J.-F. Lalonde, "Real-Time High Resolution 3D Data on the HoloLens," in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2016, pp. 189–191.

[15] H. Chen, A. S. Lee, M. Swift, and J. C. Tang, "3D Collaboration Method over HoloLens™ and Skype™ End Points," in *Proceedings of the 3rd International Workshop on Immersive Media Experiences - ImmersiveME '15*, 2015, pp. 27–30.

[16] E. Ackerman, "Google gets in your face: Google Glass offers a slightly augmented version of reality," *IEEE Spectr.*, vol. 50, no. 1, pp. 26–29, 2013.

[17] A. Hayes, "My Journey into Glass: Talking about Google Glass with stakeholders in the Glass Explorer Program," *IEEE Consum. Electron. Mag.*, vol. 5, no. 1, pp. 102–105, 2016.

[18] B. Fox and B. Felkey, "Potential Uses of Google Glass in the Pharmacy," *Hosp. Pharm.*, vol. 48, no. 9, pp. 783–784, 2013.

[19] O. J. Muensterer, M. Lacher, C. Zoeller, M. Bronstein, and J. Kübler, "Google Glass in pediatric surgery: An exploratory study," *Int. J. Surg.*, vol. 12, no. 4, pp. 281–289, 2014.

[20] B. Elise, "Google Glass a Game-changing Application in the Realm of Cultural Tourism," *Business Wire*, 2014.

[21] M. H. Schreinemacher, M. Graafland, and M. P. Schijven, "Google glass in surgery," *Surgical Innovation*, vol. 21, no. 6. pp. 651–652, 2014.

[22] A. Cave, "Why Google Glass Flopped," *Forbes*, 2015. [Online]. Available: http://www.forbes.com/sites/andrewcave/2015/01/20/a-failure-of-leadership-or-design-why-google-glass-flopped/#24c650d3556a.

[23] P. M. Corcoran, "Biometrics and Consumer Electronics: A Brave New World or the Road to Dystopia?," *IEEE Consum. Electron. Mag.*, vol. 2, no. 2, pp. 22–33, 2013.

[24] P. Corcoran, P. Bigioi, and S. Thavalengal, "Feasibility and design considerations for an iris acquisition system for smartphones," in *IEEE International Conference on Consumer Electronics - Berlin, ICCE-Berlin*, 2015, vol. 2015–Febru, no. February, pp. 164–167.

[25] S. Thavalengal, I. Andorko, A. Drimbarean, P. Bigioi, and P. Corcoran, "Proof-of-concept and evaluation of a dual function visible/NIR camera for iris authentication in smartphones," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 137–143, May 2015.

[26] S. Thavalengal, P. Bigioi, and P. Corcoran, "Iris authentication in handheld devices-considerations for constraint-free acquisition," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 245–253, May 2015.

[27] S. Thavalengal and P. Corcoran, "User Authentication on Smartphones: Focusing on iris biometrics.," *IEEE Consum. Electron. Mag.*, vol. 5, no. 2, pp. 87–93, 2016.

[28] S. Thavalengal, P. Bigioi, and P. Corcoran, "Efficient segmentation for multi-frame iris acquisition on smartphones," in *2016 IEEE International Conference on Consumer Electronics, ICCE 2016*, 2016, pp. 198–199.

[29] S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An End to End Deep Neural Network for Iris Segmentation in Unconstraint Scenarios," *arXiv Prepr. arXiv1712.02877*, 2017.

[30] H. Hofbauer, F. Alonso-Fernandez, J. Bigun, and A. Uhl, "Experimental analysis regarding the influence of iris segmentation on the recognition rate," *IET Biometrics*, vol. 5, no. 3, pp. 200–211, 2016.

[31] H. Proença and L. A. Alexandre, "Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage," *Image Vis. Comput.*, vol. 28, no. 1, pp. 202–206, 2010.

[32] M. Erbilek, M. C. Da Costa-Abreu, and M. Fairhurst, "Optimal configuration strategies for iris recognition processing," 2012.

[33] Z. Hammal, C. Massot, G. Bedoya, and A. Caplier, "Eyes Segmentation Applied to Gaze Direction and Vigilance Estimation BT - Pattern Recognition and Image Analysis," 2005, pp. 236–246.

[34] S. Natsume, "Virtual reality glasses." Google Patents, 29-Aug-2017.

[35] "CASIA Iris Image Database." [Online]. Available: http://biometrics.idealtest.org/.

[36] S. Rakshit, "Novel methods for accurate human iris recognition," *Univ. Bath*, 2007.

[37] S. Bazrafkan, T. Nedelcu, P. Filipczuk, and P. Corcoran, "Deep learning for facial expression recognition: A step closer to a smartphone that knows your moods," in *2017 IEEE International Conference on Consumer Electronics (ICCE)*, 2017, pp. 217–220.

[38] S. Thavalengal, P. Bigioi, and P. Corcoran, "Evaluation of combined visible/NIR camera for iris authentication on smartphones," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 42–49.

Appendix C

Generative Augmented Dataset and Annotation Frameworks for Artificial Intelligence (GADAFAI)

Generative Augmented Dataset and Annotation Frameworks for Artificial Intelligence (GADAFAI)

Peter Corcoran
School of Engineeing & Informatics
National University of Ireland Galway
Galway, Ireland
peter.corcoran@nuigalway.ie

Hossein Javidnia
SFI Adapt Centre,
Trinity College Dublin
Dublin, Ireland
hossein.javidnia@adaptcentre.ie

Joseph E. Lemley
Xperi Corporation.
Galway, Ireland
joe.lemley@xperi.com

Viktor Varkarakis
School of Engineering & Informatics
National University of Ireland Galway
Galway, Ireland
v.varkarakis1@nuigalway.ie

Abstract— **Recent Advances in Artificial Intelligence (AI), particularly in the field of compute vision, have been driven by the availability of large public datasets. However, as AI begins to move into embedded devices there will be a growing need for tools to acquire and re-acquire datasets from specific sensing systems to train new device models. In this paper, a roadmap in introduced for a data-acquisition framework that can build the large synthetic datasets required to train AI systems from small seed datasets. A key element to justify such a framework is the validation of the generated dataset and example results are shown from preliminary work on biometric (facial) datasets.**

Keywords—Generative Models, Data Augmentation, Data Annotation, Synthetic Data,

I. Introduction

Recent progress in Artificial Intelligence (AI) has been driven by a combination of, firstly, increased computational power, enabling more sophisticated neural network architectures to be trained; secondly the availability of large digital datasets, harvested from Big Data on the cloud and thirdly, machine learning methodologies based on deep neural networks. We are now on the cusp of a new era where embedded devices will be able to employ hardware AI accelerators [1] enabling transformative AI solutions in fields such as autonomous vehicles, smart-healthcare for the elderly, smart-city infrastructure and emerging mixed reality and wearable technologies.

But in order to realise solutions in many of these specific problem domains there is a growing need to build custom datasets that are tailored for a particular use case. Practical solutions require appropriate training datasets modified to a constrained use case together with matching ground truth data. Acquiring such datasets at the scale required for training with today's AI systems and subsequently annotating them with an accurate ground truth is challenging in terms of time, human resources and operational costs. And with the recent introduction of GDPR and associated complications introduced industry now faces additional challenges in the collection of training data that is linked to individual persons.

A. The GADAFAI Hypothesis

Recently, innovative deep learning methodologies have proved surprisingly sophisticated at generating new data samples [2]–[5] and the State-of-Art (SoA) in Virtual Reality (VR) enables large-scale photorealistic, yet virtual, 'scenes' to be created. These development suggest that it might make sense to focus on developing improved methodologies to control and manage the generation of data samples matched to a specific machine learning problem rather than struggle with the challenges of obtaining sufficient 'real-world' data. This train of thought has led to a somewhat contrarian, research question *(hereafter referred to as the "GADAFAI hypothesis")*:

"Can we artificially generate and/or augment suitably large sets of data samples adapted for training today's AI networks, and prove that the resulting AI networks are as robust and reliable as those trained on equivalent 'real-world' datasets?"

This paper outlines how GADAFAI can work in practice and provides a roadmap towards a broader validation of the hypothesis and establishes the first steps to validate the hypothesis for some specific types of dataset. To provide a context for this roadmap we focus on a range of topical research fields in Computer Vision (CV). These include, in order of increasing complexity, systems for biometric authentication (use cases: consumer devices, security & authentication), indoor scene analysis (use cases: consumer devices & home healthcare), human body motion & facial emotion analysis (use cases: home services, healthcare & security), and street-scene analysis (use cases: autonomous vehicles & smart-city). Some initial results on the validation of the GADAFAI hypothesis in the context of large facial datasets (biometrics) are provided.

B. Strategic Importance of this Research

Today AI resides mostly in cloud systems but, a migration from the datacentre towards the data sources/sensors has begun. Ultimately, a fusion of sensing & primary data-processing is to be expected on embedded devices at the network edge [6] and for some applications this is feasible today [1], [7]–[9]. But moving the core AI functions onto embedded devices is only one piece of the puzzle to deliver practical AI solutions. Currently, AI is mainly trained on Big Data, harvested from large Cloud repositories. But there are growing concerns among government regulators and the public regarding privacy issues arising from the mass harvesting of personal data. Now GDPR data privacy regulations in Europe have effectively throttled Big Data as a source of training data [10]–[12].

With many companies releasing new embedded AI accelerators [13]–[15] the only remaining barrier to new *Edge-AI* applications is training of the neural networks. But this relies on the availability of suitably annotated datasets that in most cases do not yet exist. If the full potential of *Edge-AI* [16] is to be realized in our daily lives, it is important that new

approaches are explored to building the training datasets that will enable tomorrow's disruptive AI technologies.

Currently, one of the most important challenges in building reference datasets for AI applications is the management of privacy and ethics. The EU has recently initiated major initiatives to boost research activity and encourage world-leading research on the next generation of AI in Europe [17], [18], but, in parallel, the increasing regulation of digital data poses a significant challenge to EU aspirations to become a global leader in AI research. In brief GDPR creates a significant barrier to building the large datasets required as a foundation for future AI research activity.

C. Overview of this paper

In the next section some of the challenges in acquiring large reference datasets of 'real-world' data are discussed. This is followed by a high level review of current state-of-art in data generation and augmentation techniques – these provide a foundation for GADAFAI frameworks to build large synthetic datasets. This is followed by a discussion of the primary research domains and associated data landscapes in the context of today's computer vision research. Next we outline an initial approach to validating the GADAFAI hypothesis in the context of biometric data. In this field there are large publicly available datasets and a number of well-developed methodologies to generate synthetic facial and other biometric data. A number of experiments are currently in progress to realise this initial validation and some results will be presented at the ISSC 2020 conference.

II. THE CHALLENGES OF 'REAL-WORLD' DATA

The technical challenges in building new 'real-world' datasets arise from the need for accurate annotation and expert curation of the data. If data is not correctly annotated or sufficiently generalized for the problem at hand then incorrect features are easily learned by the neural network [19]–[23].
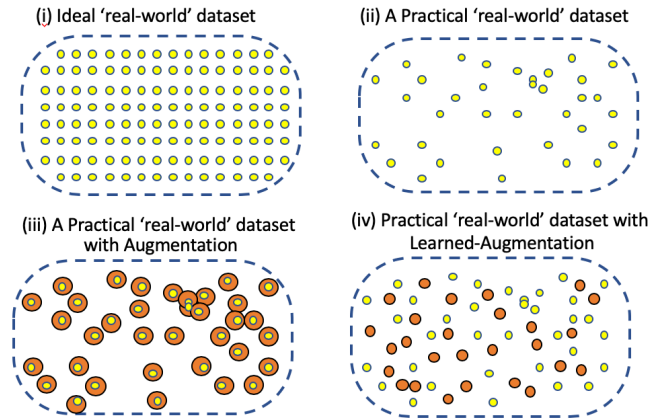

Figure 1: *Illustrative examples of datasets based on 'real-world' computer vision data.*

In *Fig. 1* some examples of datasets based on acquired 'real-world' data are shown. The ideal dataset covers all the likely sensing and environmental contexts that might arise, but as it is impractical to gather such a broadly scoped dataset researchers must typically rely on a much sparser, although still costly dataset. In practice this sparser dataset can be enhanced using simple data augmentation techniques [24], [25] to grow the context of the original samples. More sophisticated technique such as learned augmentation [26], [27] can help to create some new, intermediate data samples. This can improve the performance of existing datasets, but it still relies on a large original 'real-world' dataset which is

costly and time-consuming to acquire and often provides a less that reliable ground truth.

To better illustrate the challenges of real-world data it is useful to consider a simple example – acquiring a stereo baseline dataset for training a neural network to detect image depth. To acquire this dataset, a researcher needs two identical cameras – but individual cameras have subtle differences in the optics (distortions & aberrations), the CMOS sensors (noise) and in the digital pipelines (jitter & rolling shutter synch). To obtain well matched stereo pairs a detailed baseline calibration should be performed between the two cameras. Subsequently, they should be maintained at the same temperature and in similar ambient conditions. An accurate depth sensing system is also required as part of the experimental rig to gather detailed depth information for a ground truth; the point cloud of depth measurements needs to be aligned with individual pixels of the camera image for both cameras, and for each image frame as the imaged scene changes. Finally, if it subsequently becomes necessary to change the stereo baseline, the sequence of imaged scenes must be re-created frame by frame, and if the original data is to be re-used then each scene should be re-created exactly as the original (lighting, object poses & locations, stereo path, etc). This exact re-creation of the experimental conditions is practically impossible to achieve in the 'real-world'.

Now consider an implementation of the above acquisition sequence in a photo-realistic virtual environment. The two cameras are represented by 'camera models' and are 'identical' digital twins of each other; the exact depth values for each rendered pixel are available and the stereo acquisition paths are precisely defined and can be readily adjusted. Objects and animations in the 3D scene are stored and can be recreated and manipulated at will in a repeatable manner.

Naturally the storage and rendering of complex 3D scenes has significant associated computational and storage overheads so it is helpful if we can also leverage advanced augmentation techniques to grow the size of the rendered dataset. Further, 3D scenes typically employ digitally generated components which do not provide the same variations as 'real-world' objects and textures – thus some acquisition of 'seed data' from the real world is still important to better generalize the training data and introduce realistic variations in surface texture, lighting and image noise.

III. STATE OF ART IN DATA GENERATION & AUGMENTATION

A. Advanced Data Augmentation

Fully learnable data augmentation was originally proposed by Lemley et al. [26] and there have been many refinements of the approach over the last 3 years [27]. In the learned, or smart augmentation technique of [26], all the components of the augmentation pipeline are learned via an auxiliary neural network. An alternative approach is proposed in [28] where the augmenter accepts two random images from a class and tries to generate images which reduce the loss of the target network (a network that learns a desired task). In another approach a Bayesian technique [29] is applied to generate data based on the distribution which is learned from an originl training set. Similarly, learned features can be manipulated using simple transformation which results in augmented data as shown in [30]. Note that learned augmentation techniques differ from the more widely researched generative adversarial networks (GANs) in that the end point is create data samples that improve a training task, rather than data samples that attempt to mimic a known data distribution.

B. Generative Adversarial Networks (GANs)

Since the introduction of GANs in 2014 [31], there has been a rapid growth of research in this field. GANs are now applied to many applications across different domains from computer vision to natural language processing and audio/speech processing. The use of GANs in SoA is particularly effective in the generation of facial data samples [32]–[36]. The GADAFAI hypothesis seeks to investigate a broader range of data classes across the image and audio domain, but initial investigations have focussed on facial data as advanced techniques and large public datasets are more highly evolved in this field.

IV. OVERVIEW OF GADAFAI

GADAFAI is distinguished from other efforts to build synthetic datasets in that it also seeks to provide some validation metrics to measure the usefulness of the synthetic data samples for a particular use case or problem. These metrics should allow researchers to quantify variances between 'real-world' datasets and those that are 'generated' via alternative methodological approaches. The intention is to focus on measuring the validity of synthetic datasets for practical problems across a range of fields of application. This should enable new methodological refinements of the generated datasets to specific use cases.

Key research fields to be addressed include, in order of increasing complexity and challenge, (i) biometrics (consumer devices, security & authentication); (ii) indoor scenes (smart-home & home healthcare), (iii) human body motion & facial emotion analysis (home services, healthcare & security), (iv) complex & dynamic street scenes (autonomous driving & smart-city) and (v) human speech synthesis. Each of these research fields provide a more complex format of dataset and represent progressive challenges both in generating appropriate and accurate datasets and in the availability of suitable quantitative and qualitative metrics to realize the validation of the dataset.

A. The Data Landscape

There is no doubt that we face a complex and heterogeneous data landscape in the field of computer vision and much of the challenge revolves around this complex data landscape. The GADAFAI hypothesis considers this data in a series of categories of increasing complexity. Note that the main goal is to demonstrate sufficient commonality to deliver on a broadly-scoped data synthesis framework, while potentially solving some specific applied research problems. Here we briefly review each data category, summarising the inherent challenges.

Biometric Data – This category of data is well studied in the literature and mature techniques to analyse and validate biometric identity are available, especially for facial and iris recognition pipelines. More recently, advanced deep learning techniques have shown improvements on certain aspects of these acquisition pipeline and the use of less robust acquisition platforms, such as handheld devices have been explored. The maturity of biometric data analysis offers a good starting point to explore the GADAFAI hypothesis and some work on the validation of synthetic facial data samples for use in the area of facial recognition will soon be ready for publication.

Indoor Scenes – This category of data has received much attention recently as methods for acquiring 2.5D data (imaged scene + depth data) have been refined and in parallel there have been advances in the 3D modelling of rooms and buildings and recently a range of deep learning techniques and the use of GANs have been applied to both analyse indoor scenes and generate random, but semantically valid models of indoor spaces. Important tools include monocular depth estimation [37], [38] and scene segmentation [39]. This data category provides a first step away from the maturity of biometric data, but as scenes are essentially static the challenges are reasonably tractable. Applied problems related to this data category include the dynamic analysis of a room environment to improve consumer audio experience, or to enable augmented or mixed reality on smartphones or with wearable glasses. This data category also offers a stepping-stone to modelling a dynamic indoor environment involving human interactions.

Body Motion & Facial Expressions – This spans a number of different areas from motion-capture, typically for purposes of advanced animations, to security systems, often in urban environments and more recently Driver-Monitoring Systems (DMS) and Adaptive Driver Assistance Systems (ADAS) for the automotive sector. This data category presents an additional challenge as it involves dynamic transitions rather than the static data of the first two categories. Research is not as mature, but there are many useful datasets and data analysis tools that can be leveraged to support data-generation, augmentation and 3D modelling.

Indoor Scenes with Humans – This data category takes us beyond current state-of-art by combining the previous two data categories to enable the modelling or generation of dynamic human interactions within a living environment. At this point technical challenges are encountered in terms of the size of data required to represent such dynamic scenes and the bandwidth and computational power required to model or generate interactive action sequences. Some real-world datasets exist for human activities and actions, but they are less comprehensive than in the previous data categories. Applied problems related to this category include monitoring the capacity of the stay-at-home elderly – a compelling and global socio-economic challenge.

Street Scenes with Dynamic Activities – This data category involves a wider range of depths and more complex dynamic scenes with multiple vehicles, humans and other animated elements. While these scenes represent an even greater challenge over dynamic indoor scenes some aspects of the data models and generative techniques are better developed in this field due to the research focus of industry and academia on autonomous driving.

B. The GADAFAI Data-Generation Framework

Up to this point the goal was to explain the scope and context of the GADAFAI hypothesis. Here the proposed dataset-generation framework is explained in the context of a practical application such as generating a dataset of facial identities to train a facial recognition classifier. *Figure 2* illustrates a brief overview of the pipeline employed to activate GADAFAI hypothesis. The process starts with a number of seed datasets - Fig 2(i). For facial data samples there are quite a few suitable datasets, one of the best being CELEB-A. Public facial datasets typically have a significant number of bad data samples. These may be incorrectly labelled or of such poor quality that a neural network algorithm will not be able to learn correctly from these samples. Thus the seed dataset often has to be pre-filtered, or cleaned before if can be used. Another challenge is that the number of data samples in each class can vary widely with

some classes only having 1-3 data samples. These classes should be excised from training data but can be useful later for validation. Classes with more samples, but lower numbers – typically having only 5-10 samples – should have additional data samples created using learned augmentation - Fig 2(ii).
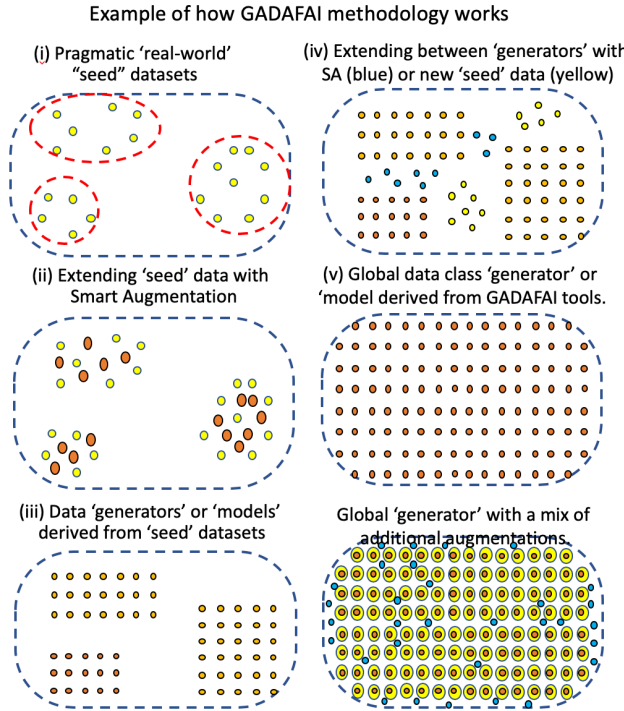

Figure 2: *The brief overview of GADAFAI hypothesis*

To create additional data classes a GAN such as StyleGAN should be trained - Fig 2(iii) – and these can then be used to generate new data classes which can bridge gaps between the seed datasets - Fig 2(iv). At this point the framework can provide a larger hybrid dataset than is available from public datasets, and such a comprehensive dataset should, on its own, enable more sophisticated facial classification techniques and building more adaptive facial class generators. Current state of research for this specific problem is now at the stage illustrated in Fig 2(iv) and should be published later this year.

V. PRACTICAL VALIDATION ON FACIAL DATASETS

It is clear that, the GADAFAI hypothesis is somewhat contrarian in nature - no data scientist would subscribe to the idea that man-made data can substitute for real-world data. Thus, it is important to validate the hypothesis across a number of different fields. As was discussed earlier, the field of facial biometrics offers a good starting point. In this section some details of our initial work on facial datasets is provided, together with some preliminary outcomes.

There are several steps to achieve this first validation. Firstly it is necessary to build tools that can generate a larger number of unique synthetic data samples of a human face. Secondly we need to work with existing 'real-world' facial datasets to obtain a relatively clean distributions of data samples. Both of these tasks are challenging in their own right and are documented in companion papers to be presented at ISSC 2020 and QoMEx 2020. A third, more comprehensive journal paper will bring together all of this research. In this section an overview of the current status and summary of the main findings is presented.

A. Re-Training of StyleGAN

StyleGAN is a state-of-art generative adversarial network (GAN) architecture that generates random 2D high-quality synthetic facial data samples. The original GAN network for StyleGAN was trained on the CelebA-HQ (high quality) dataset which has 30,000 facial images. However this is not a particularly large facial dataset so the motivation to re-train StyleGAN is to better understand the relationship between the original seed data and the synthetic samples generated by the GAN after training on different sizes of seed data.
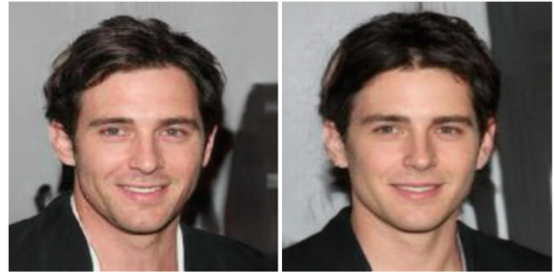
Figure 3: *Four examples of uncannily similar StyleGAN image pairs generated randomly in a batch sample of 20,000 images;*

Our preliminary work suggests that without a sufficiently large seed dataset that StyleGAN can exhibit overlappings within the generated samples within the latent space. This manifests through a small number of synthetic data samples that are separated in latent space, but appear uncannily similar to one another as shown in *Figure 3*. Additional experiments are in progress to better quantify this phenomenon. The results of several of our re-trainings of the generator network are

made available publicly [1] to encourage other researchers to contribute and explore the stability of StyleGAN further.

B. Cleaning of Large Facial Datasets

CelebA and CelebA-HQ are actually not very optimal datasets as they provide relatively small numbers of individual data samples and separate identities. CelebA has 200,000+ individual facial data samples and 10,000+ identities, but that is only an average of 20 samples per identity. There are several datasets with at least an order of magnitude more unique data samples and thus a greater density of samples per identity. But all these large datasets share one common challenge – they have many bad data samples. Some are of poor image quality, but there are also many examples of bad labelling of data in all public datasets. This motivated a second piece of work to find a method to improve the quality of such large real-world datasets because without such clean 'real-world' datasets we cannot make a fair comparison with datasets built from synthetic data samples.

C. Real Vs Synthetic Data Samples - Preliminary ROCs

Our initial experiments comparing CelebA negative pairs with StyleGAN and several other datasets led to the discovery of overlapping identities across several of these datasets. After removal of these overlapping identities the results shown in Figure 4. It is noted that all of the ROC curves are essentially consistent, apart from a comparison of StyleGAN with StyleGAN negative pairs (yellow curve). At this point it was concluded that this poor performance is due to the similar negative pairs, as illustrated in *Figure 3* and this is turn is due to insufficient training samples in the CelebA dataset. A re-training of StyleGAN on the larger CelebA and Casia datasets was recently completed and results from an extended set of experiments will be provided at the ISSC 2020 conference.
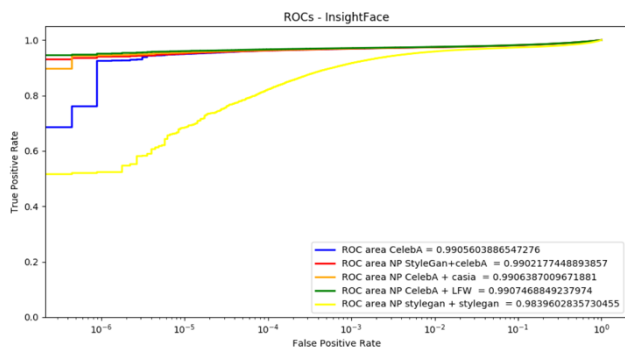

Figure 4*: ROC curves comparing negative pairs between CelebA and StyleGAN, Casia and Labeled Faces in the Wild (LFW) datasets.*

D. Towards complete Synthetic Identities

The results presented here are the first phase of validation on large facial datasets. To complete the validation the next step is to build a complete facial identity from a random StyleGAN data sample so that a fair comparison can be made between a 'clean' real-world dataset and a corresponding 'synthetic' dataset. This entails augmenting each random StyleGAN data sample with variations such as facial expressions, pose variations, and mixed lighting conditions amongst others. These variations can be generated both using conventional image processing techniques and though specialized training of GANs. Elements of this work are underway and will be reported in detail in a following publication.

VI. DISCUSSION

The initial work on validation of the GADAFAI hypothesis has led to some interesting findings and outputs. Firstly, in order to provide clean versions of a number of large public face datasets a semi-automated cleaning methodology was developed and validated and will be presented in a companion paper.

Secondly, the quality of the data samples generated by StyleGAN when trained on some of the smaller facial datasets, such as CelebA-HQ or CELEBA has been explored. Initial results show that some of the generated data samples result in uncannily similar facial identities. This suggests that either there is insufficient variation in the seed data used to train the generator or alternatively the GAN may be approaching mode collapse. More detailed investigations are ongoing and a methodology to measure the uniqueness of generated data samples is evolving. These results should inform other researchers relying on the use of generators such as StyleGAN that additional analysis of the generated data is needed if this data is to be used to accurately simulate variations in real-world datasets. Some additional results will be presented at the conference and a more detailed journal publication is in preparation.

The GADAFAI research hypothesis has been outlined in some detail in this work and represents an evolving approach that can help develop powerful new training methodologies to enhance the capabilities of state-of-art neural networks. Ultimately, such methodological frameworks could free researchers from concerns with the logistics of dataset acquisition, enabling them to focus on new technology innovations in terms of smart services and products.

REFERENCES

[1] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision.," *IEEE Consum. Electron. Mag.*, vol. 6, no. 2, pp. 48–56, Apr. 2017.

[2] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Networks," *arXiv preprint arXiv: ...*, 2014. [Online]. Available: http://arxiv.org/abs/1406.2661.

[3] Yoshua Bengio, C. Li, B. Zhang, and T. Shi, "Deep Generative Models," *Deep Learn.*, pp. 658–729, 2016.

[4] S. Bazrafkan, H. Javidnia, P. C. preprint arXiv:1802.00390, and undefined 2018, "Face Synthesis with Landmark Points from Generative Adversarial Networks and Inverse Latent Space Mapping," *arxiv.org*.

[5] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[6] P. Corcoran, "Mobile-Edge Computing and Internet of Things for Consumers: Part II: Energy efficiency, connectivity, and economic development," *IEEE Consum. Electron. Mag.*, 2017.

[7] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Efficient CNN Implementation for Eye-Gaze Estimation on Low-Power/Low-Quality Consumer Imaging Systems," *arXiv preprint arXiv:1806.10890*, 2018.

[8] S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end Deep Neural Network for iris segmentation in unconstrained scenarios," *Neural Networks*, vol. 106, pp. 79–95, Oct. 2018.

[9] V. Varkarakis, S. Bazrafkan, and P. Corcoran, "A Deep Learning Approach to Segmentation of Distorted Iris Regions in Head-Mounted Displays," in *2018 IEEE Games, Entertainment, Media Conference*, 2018, pp. 1–9.

[10] T. Z. Zarsky, "Incompatible: The GDPR in the Age of Big Data," *Iowa Law Rev.*, 2017.

---

[1] https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Re-training-StyleGAN

[11] M. Butterworth, "The ICO and artificial intelligence: The role of fairness in the GDPR framework," *Comput. Law Secur. Rev.*, 2018.

[12] V. Mayer-Schönberger and Y. Padova, "REGIME CHANGE? ENABLING BIG DATA THROUGH EUROPE'S NEW DATA PROTECTION REGULATION," *T H E C O L U M B I A Sci. Technol. LAW Rev.*, 2016.

[13] K. Kwon, A. Amid, A. Gholami, B. Wu, K. Asanovic, and K. Keutzer, "Invited: Co-Design of Deep Neural Nets and Neural Net Accelerators for Embedded Vision Applications," in *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1–6.

[14] H. Yoo, "1.2 Intelligence on Silicon: From Deep-Neural-Network Accelerators to Brain Mimicking AI-SoCs," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 20–26.

[15] T. Fujii *et al.*, "New Generation Dynamically Reconfigurable Processor Technology for Accelerating Embedded AI Applications," in *2018 IEEE Symposium on VLSI Circuits*, 2018, pp. 41–42.

[16] P. Corcoran, J. Lemley, C. Costache, and V. Varkarakis, "Deep Learning for Consumer Devices and Services 2—AI Gets Embedded at the Edge," *IEEE Consum. Electron. Mag.*, vol. 8, no. 5, pp. 10–19, Sep. 2019.

[17] C. Huet, "European Commission's Initiatives in Artificial Intelligence," 2017. [Online]. Available: https://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-huet.pdf. [Accessed: 05-Jul-2018].

[18] "Twenty-four EU countries sign artificial intelligence pact in bid to compete with US and China · AI-Hub Europe." [Online]. Available: http://ai-europe.eu/twenty-four-eu-countries-sign-artificial-intelligence-pact-in-bid-to-compete-with-us-and-china/. [Accessed: 05-Jul-2018].

[19] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," *CVPR*, pp. 2574–2582, 2016.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arxiv.org*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6572. [Accessed: 28-Jun-2018].

[21] C. Szegedy *et al.*, "Intriguing properties of neural networks," 20-Dec-2013. [Online]. Available: http://arxiv.org/abs/1312.6199. [Accessed: 28-Jun-2018].

[22] J. Jin, A. Dundar, and E. Culurciello, "Robust Convolutional Neural Networks under Adversarial Noise," Nov. 2015.

[23] R. D. Hjelm, A. P. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio, "Boundary-seeking Generative Adversarial Networks," in *ICLR*, 2018, pp. 1–17.

[24] P. Corcoran, C. Costacke, V. Varkarakis, and J. Lemley, "Deep Learning for Consumer Devices and Services 3-Getting More from Your Datasets with Data Augmentation," *IEEE Consum. Electron. Mag.*, 2020.

[25] P. Corcoran, J. Lemley, and S. Bazrafkan, "Getting more from your datasets: Data augmentation, annotation and generative techniques," in *Embedded Vision Summit*, 2018.

[26] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[27] J. Lemley and P. Corcoran, "Deep Learning for Consumer Devices and Services 4-A Review of Learnable Data Augmentation Strategies for Improved Training of Deep Neural Networks," *IEEE Consum. Electron. Mag.*, 2020.

[28] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv Prepr. arXiv1712.04621*, 2017.

[29] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A bayesian data augmentation approach for learning deep models," in *Advances in neural information processing systems*, 2017, pp. 2797–2806.

[30] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," *arXiv Prepr. arXiv1702.05538*, 2017.

[31] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Networks," *arXiv preprint arXiv:1406.2661*, 2014. .

[32] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv Prepr. arXiv1703.10717*, 2017.

[33] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv Prepr. arXiv1710.10196*, 2017.

[34] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[35] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10215–10224.

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv Prepr. arXiv1701.07875*, 2017.

[37] S. Bazrafkan, H. Javidnia, and J. Lemley, "Semiparallel deep neural network hybrid architecture: first application on depth from monocular camera," *J. Electron. Imaging*, 2018.

[38] F. Khan, S. Salahuddin, and H. Javidnia, "Deep learning-based monocular depth estimation methods—a state-of-the-art review," *Sensors (Switzerland)*. 2020.

[39] M. Naseer, S. H. Khan, and F. Porikli, "Indoor Scene Understanding in 2.5/3D: A Survey," *IEEE Access*, 2018.

Appendix D

Re-Training StyleGAN-A First Step Towards Building Large Scalable Synthetic Facial Datasets

Re-Training StyleGAN - A First Step Towards Building Large, Scalable Synthetic Facial Datasets

Viktor Varkarakis
Collage of Engineering and Informatics
National University of Ireland, Galway
Galway, Ireland
v.varkarakis1@nuigalway.ie

Shabab Bazrafkan
Imec VisionLab, Physics Department
University of Antwerp
Antwerp, Belgium
shabab.bazrafkan@uantwerpen.be

Peter Corcoran
Collage of Engineering and Informatics
National University of Ireland, Galway
Galway, Ireland
peter.corcoran@nuigalway.ie

Abstract—**StyleGAN is a state-of-art generative adversarial network architecture that generates random 2D high-quality synthetic facial data samples. In this paper we recap the StyleGAN architecture and training methodology and present our experiences of retraining it on a number of alternative public datasets. Practical issues and challenges arising from the retraining process are discussed. Tests and validation results are presented and a comparative analysis of several different re-trained StyleGAN weightings is provided. The role of this tool in building large, scalable datasets of synthetic facial data is also discussed.**

Keywords—sunthetic face data, face recognition, generative adversarial networks, GANs, StyleGAN

I. INTRODUCTION

In the last few years, a number of tools for generating synthetic facial data samples have evolved [1]–[3], based on generative adversarial networks (GANs) [4]. These enable photo-realistic, high-resolution random face samples to be generated in almost limitless numbers. However, these face samples are essentially random with no relationship to one another, and thus of limited value or interest to researchers.

Some research has attempted to modify the training of these GANs to provide more control over the output data samples. As examples, Bazrafkan et al [5] have shown that a GAN can be trained so that the first vector in the latent space represents male samples if it has a negative value or female images if the value is positive. In a separate work, Bazrafkan et al [6] have shown that faces can be generated in a fixed pose set by an auxiliary regressor.

From these works, it is clear that GANs can be adapted and modified to enable more sophisticated control of the generation of synthetic facial data samples. It is also important to understand the relationship between the original training dataset and the data samples obtained from the resulting GAN. In order to obtain a wide variation in output data, it is necessary to have a large original training dataset, but it is not clear from current research how the size and quality of this original dataset affect the distribution of the output data samples. Neither is it clear how well these synthetic data samples are distinguished from the original data.

StyleGAN has been used widely and trained on different image topics (cats, cars, bedrooms, anime, etc) [3][7]. The original models released in StyleGAN [3], provide models trained on the FFHQ [3] and CelebA-HQ [8] datasets at a resolution of 1024×1024. It is worth noting that the size of these datasets is relatively small as they consist of 70000 and 30000 samples respectively. Apart from these original models to the best of our knowledge a few models are publicly available that provide StyleGAN models trained on alternative facial datasets at different image resolutions and quality. These models are unofficial implementations [1],[2] of the StyleGAN in other frameworks (PyTorch) and therefore it is not possible to compare them with the original StyleGAN [3].

As a first step to understanding better the nature of state-of-art GANs we have re-trained StyleGAN on a number of large publicly available datasets and made the resulting model networks with their weights publicly available [3]. In this paper, we describe the re-training process on various original facial datasets and explain the various steps in the re-training process. We then perform a comparative analysis of a randomized set of data created by each of the re-trained GANs, in order to evaluate the quality and diversity of the generated samples.

II. AN OVERVIEW OF PUBLICLY AVAIALBLE FACIAL DATASETS

In recent years more and more "in the wild" face datasets have become available of different sizes and with each dataset being proposed for a different use case. A short description of some of the more significant publicly available facial datasets is given below.

A. LFW

Labelled Faces in the Wild (LFW) [9] 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

---

[1] https://github.com/kayoyin/FakeCelebs
[2] https://github.com/podgorskiy/StyleGan
[3] https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Re-training-StyleGAN

LFW, it is inadequate for training purposes and thus is used mainly for testing.

B. CASIA-WebFace

CASIA-WebFace [10] 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.

C. CelebFaces

The CelebFaces+ dataset [11] 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) [12] where the samples form CelebFaces+ are annotated with 5 landmark locations and for 40 binary attributes (eyeglasses, moustache, hat, etc), providing valuable information for the researchers.

D. MegaFace

The MegaFace dataset [13] was published in 2016 in order to examine face recognition methods with up to a million distractors in the gallery image set. The dataset consists of 4.7M samples organised into 672,057 identities. Despite being a large dataset, it offers a limited set of variations per identity, as on average it has only 7 samples per person. Depending on the specific research goal this can limit the usefulness of MegaFace for some research tasks.

E. Ms-Celeb-1M

The Ms-Celeb-1M dataset [14] was created and published in 2016 by Microsoft. It is the largest publicly available face recognition dataset with over 10M samples from 100K identities. The dataset is suitable for both training and testing purposes with an average of 100 samples per identity.

F. VGGFace & VGGFace2

The VGG datasets are released from the Visual Geometry Group from the University of Oxford. The VGGFace [15] dataset was released in 2015 and contains 2.6M samples from 2,622 people. VGGFace was released mainly for training purposes. In 2018, the VGGFace2 [16] was released. This dataset comprises 3.31M samples from 9,131 celebrities – on average 360 samples per identity. The images were downloaded from Google Image Search. The image samples from VGGFace2 cover a wider range of different ethnicities, professions, and ages compared to VGGFace. Furthermore, all the samples have been captured "in the wild" thus giving the dataset a desirable variation with respect to pose, lighting and occlusion conditions as well as facial expressions. The dataset can be used for training and testing purposes as it is divided into a train and test set. Finally, VGGFace2 provides annotations regarding the pose and the age of its samples which can be useful for researchers.

G. Other Face Datasets

Several other datasets should be mentioned such as YTF [17], which has 1,595 identities and 3,425 video clips. Another dataset that was built in order to recognise faces in unconstrained videos is UMDFaces-Videos [18], which consists of 3,107 identities and 22,075 video clips. Also, they have a face dataset as well with still images, the UMDFaces which consist of 367,88 samples from 8,277 identities [19].

Furthermore, FFHQ [3] and CelebA-HQ [8] are some face datasets that are not created for face recognition purposes. These datasets are of a high quality and a high resolution 1024×1024 compared to the aforementioned databases. These datasets were used to train StyleGAN, which produces high quality generated images. The FFHQ consists of 70,000 images without an identity annotation but contains variation in terms of age, ethnicity and image background. It also has a good coverage of accessories. CelebA-HQ is a subset of CelebA from 6,217 identities. As mentioned before the samples are at 1024×1024 resolution and of high quality which was achieved through a procedure of pre-processing that is explained in [8].

Finally, it is also worth remarking that large corporations, e.g. Facebook, Google, have their own in-house datasets that are likely to dwarf those that are publicly available. Facebook [20] trained some of their model with a dataset that comprises 500M facial samples from more than 10M identities and Google's model [21] was trained on 200M images from 8 million subjects.

III. RETRAINING METHODOLOGY

In this section, the process of retraining StyleGAN is described. Initially, a preparatory filtering procedure is used to select the samples is explained and subsequently, the procedure of training StyleGAN is analysed. For the purposes of this work, the StyleGAN network was retrained twice, once with samples from the CelebA dataset and the other one is trained on the CASIA-WebFace dataset. The techniques documented here are being applied to additional datasets and corresponding results will be presented at the ISSC conference later this year.

A. Data Sample Resolution and Quality Considerations

The datasets used in our experiments (CelebA, CASIA-WebFace) are pre-filtered before the data samples are fed into the training network. The filtering is performed for two main reasons. Firstly, it is important to ensure that the data samples given to the network contain a detectable face region of good quality. Most large face datasets contain noisy, poor samples and it is desirable to remove such samples as inputs to the training network as they can interfere with the main learning task of generating realistic face samples of good visual quality. Secondly, it is important to resize the facial samples to a particular size that the StyleGAN network will be trained on. The size of the image samples used in this work was determined to be 256×256 pixels. This offers a good balance between facial image quality and the computational resources required for training. With such a size of image samples, it is practical to train on a single dual-GPUs computer. Based on the

information available on StyleGAN's GitHub repository [4], training at a higher resolution such as 1024×1024 with only two GPUs it would have required almost a month of continuous operation.

For our purposes – to gain practical experience in re-training these powerful GANs - samples at a resolution of 256×256 have sufficient visual information and quality for most practical uses and applications of synthetic facial data. It is also worth noting that most of the available public face datasets the data samples are of similar or lower resolution and often there is a significant number of samples that are noisy or of quite poor visual quality.

B. Dataset Preparation & Pre-Filtering

In the initial filtering step, the image samples are passed through a face detector. The face detector used is the Multi-task Cascaded Convolutional Networks (MTCNN) [22]. An implementation of the MTCNN in Python / TensorFlow was used which can be found in [5]. The face detector is applied to the images. The MTCNN implementation used takes two arguments, the margin size and the size of the output image. The margin used is 50px and the size of the image as mentioned earlier is 256×256. The CelebA dataset has 202,599 samples and the CASIA-WebFace has 494,414 samples. After the pre-processing procedure, CelebA and CASIA-WebFace consist of 202,281 and 491,073 samples respectively. In Fig 1. and Fig 2. some samples from CASIA-WebFace and CelebA are presented. The MTCNN detection was not able to confirm these samples as faces and thus they are not used in training. These examples illustrate the need for a pre-filtering step for the input data. In all large public face datasets, a significant number of such noisy data samples are expected and may unduly affect the training outcome. In Fig. 1 there are some examples of extreme pose, images with artifacts, blurred or extremely dark facial images or only partial face samples. In Fig. 2, samples are presented that only contain noise, or the face is mostly obscured and is not representative of a normal human face.

With the use of MTCNN to pre-filter the data, it is possible to a certain degree to eliminate many unwanted samples that would not be beneficial for retraining StyleGAN for the task of generating unobscured, human faces. Finally, in Fig. 3 and Fig. 4 a selection of good, high-quality, facial samples from CASIA-WebFace and CelebA are presented after the pre-processing procedure and used to prepare data samples for the main training procedure.

Fig. 1. CASIA-WebFace samples [10], not detected by the MTCNN [22].

Fig. 2. CelebA samples [12], not detected by the MTCNN [22].

Fig. 3. CASIA-WebFace samples [10], after the pre-processing procedure

Fig. 4 CelebA samples [12] after the pre-processing procedure.

C. The Re-Training Process

The original Generative Adversarial Networks (GAN) presented in [4] is made of two Deep Neural Networks: a generator and a discriminator. The generator accepts a tensor of randomly generated numbers and returns an image and the discriminator is a binary classifier that accepts an image and determines whether it is a generated image or not. In this approach, these two networks are

---

[4] https://github.com/NVlabs/stylegan

[5] https://github.com/davidsandberg/facenet

trained in a min-max game wherein the final goal is for the generator to synthesis an image that the discriminator classifies as a real image.

The StyleGAN [3] is one of the variations of GAN wherein the generator is developed in a specific way which separates it from its preceding implementations in three main ways:

1- The latent space (Z) is reshaped via a fully connected DNN (which returns W) before feeding into the generator. This is to introduce disentanglement to the original latent space (Z) during the mapping into style indicators (W).

2- The latent space is not fed into the generator at its input layer. The new latent space W is given to the generator before each convolutional layer. In other words, each part of the vector W is induced into the generator in a different layer. This gives the opportunity to introduce style information at different levels.

3- A Gaussian noise is added to the features before each convolution. This operation helps the network to use its maximum capacity and generate higher quality outputs with high-frequency features.

More details regarding the architecture as well as the hyperparameter selection of StyleGAN can be found in [3].

In this current study, once the samples were pre-processed, StyleGAN is re-trained on the two large datasets mentioned above (CelebA, CASIA-WebFace). The official implementation of StyleGAN was adopted to retrain on our databases. This implementation is in TensorFlow, requiring version 1.10 or newer and Python 3.6 and can be found in [4]. The default configuration for training was utilized which used to train the highest-quality StyleGAN with the FFHQ dataset at a 1024×1024 resolution. Full details can be found in [3], as here we only described modifications to the base training. As mentioned earlier a dual-GPU machine was used to train StyleGAN. The GPUs used are RTX 2080 Ti. The training process was performed twice, once training with samples from CelebA dataset and once with samples from the CASIA-WebFace dataset. Each experiment ran for 12 days. After the end of the training process, the epoch/checkpoint with the best Fréchet Inception Distance [23] using 50,000 images (FID50k) is selected. FID50k is an evaluation metric used in the training procedure. In the next section, examples and the evaluation results for the best models are presented. Finally, we make these models publicly available and can be found in [3].

IV. RETRAINING – EXPERIMENTS & VALIDATION

As mentioned, the model with the best Fréchet inception distance (FID) using 50,000 images for each dataset with which the StyleGAN was trained (CelebA, CASIA-WebFace) is selected as the final model. The FID is an evaluation metric that captures the similarity of generated images to real ones better than the Inception Score [23]. A lower FID score means better image quality and diversity of the generated images. For the best model selected for each dataset, the results of several quality and disentanglement metrics are presented along with visual examples. More information about the metrics and the way that are calculated can be found in [3].

Table I. Quality and disentangelement metrics for several StyleGAN models trained on different datasets and resolutions. Specifically, SyleGAN trained on CelebA and Casia-WebFaces at 256 ×256 and the original StyleGAN [3] trained on FFHQ at 1024×1024.

Metric	Results			Description
	StyleGAN on CelebA	StyleGAN on CASIA-WebFace	StyleGAN on FFHQ [3]	
FID50k	4.7842	4.5992	4.4159	Fréchet Inception Distance using 50,000 images.
ppl_zfull	191.9051	258.4270	664.8854	Perceptual Path Length for full paths in Z.
ppl_wfull	68.6066	81.7605	233.3059	Perceptual Path Length for full paths in W.
ppl_zend	190.5838	259.5282	666.1057	Perceptual Path Length for path endpoints in Z.
ppl_wend	56.4555	74.2621	197.2266	Perceptual Path Length for path endpoints in W
ls in Z	143.2236	109.7136	165.0106	Linear Separability in Z
ls in W	2.5235	3.1748	3.7447	Linear Separability in W

A. Evaluation of the re-trained StyleGAN models

In Table I, the quality and disentanglement metrics for the best StyleGAN models on each dataset are presented. The selected model of StyleGAN trained on the CelebA dataset, has a better score in the perceptual path length for full paths and for the endpoints path in W and Z latent space, compared to the selected model trained on CASIA-WebFace. Perceptual path length [24] measures the difference between consecutive images (their VGG16 embeddings) when interpolating between two random inputs. Drastic changes mean that multiple features have changed together and that they might be entangled, therefore showing that the model trained with the CelebA dataset generates samples that its features are less connected between them compared to the ones from the model trained with CASIA-WebFace. The metric of linear separability shows the ability to classify inputs into binary classes. The better the classification the more separable the features. In this metric, the model trained with CelebA has a better score in the W latent space whereas the model trained on CASIA-WebFace has a better score in the Z latent space. Finally, in the FID score using 50,000 images the StyleGAN model trained on CASIA-WebFace has a better score illustrating a slightly better quality and diversity in the generated samples compared to the StyleGAN model from CelebA. In Table I, the quality and

disentanglement metrics for the StyleGAN trained on FFHQ from [3], are presented. We do not compare with their results as they trained StyleGAN on high quality and high resolution (1024×1024) in contrast with our models which they were trained on a ×4 smaller resolution and lower quality samples.

Below several samples are generated from the trained StyleGAN models. Fig. 5 and 6 show an uncurated set of novel images generated from the trained generator. Fig. 7. shows the effect of applying stochastic variation to different subsets of layers. Fig 8 illustrates the effect of the truncation trick as a function of style scale ψ. The samples in Fig 5, and 8 are generated by the selected StyleGAN model trained with CelebA and Fig 6 and 7 with the selected StyleGAN model trained with CASIA-WebFace. The models used to generate the samples are available in [3].

Fig. 5. Uncurated set of images produced by the best StyleGAN model trained with CelebA. The samples are generated with a variation of the truncation trick [25]–[27], ψ = 0.7 for resolutions $4^2 − 32^2$. This figure is similar to the figure (2) from [3]

Fig. 6. Uncurated set of images produced by the best StyleGAN model trained with CASIA-WebFace. The samples are generated with a variation of the truncation trick [25]–[27], ψ = 0.7 for resolutions $4^2 − 32^2$. This figure is similar to the figure (2) from [3].

Fig. 7. Effect of noise inputs at different layers of the generator. (a) Noise is applied to all layers. (b) No noise. (c) Noise in fine layers only ($64^2 − 1024^2$). (d) Noise in coarse layers only ($4^2 − 32^2$). This is similar to figure (5) from [3].
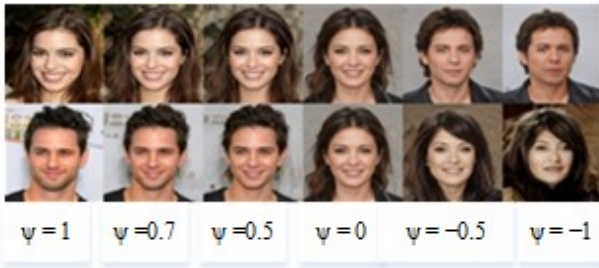
Figure 8 The effect of truncation trick as a function of style scale ψ (ψ=1). When we fade ψ → 0, all faces converge to the "mean" face of CelebA. This is similar to figure (8) from [3]

V. DISCUSSION AND FUTURE WORK

StyleGAN is currently the state-of-the-art in generating images, especially in the task of realistic face generation. In the sections above the procedure of re-training it on several public face datasets is discussed. The network was re-trained on two large publicly available datasets, CelebA and CASIA-WebFace. The Original StyleGAN has been trained on the FFHQ and CelebA-HQ face datasets which are of high quality and high resolution in [3]. As in this work, the training is performed with images of low resolution and lower quality, no comparison is being performed between the StyleGAN models of this work and the models from [3]. We trained StyleGAN models on different face datasets with different resolutions providing a useful tool for researchers, as to make these models available in ³. Furthermore, it gives the opportunity to examine several aspects of StyleGAN. As mentioned StyleGAN is being trained on other large face datasets and further results will be presented at the ISSC conference later this year. It should be noted that this work is a first step and an important tool that will be used in order to understand how the size and quality of the original dataset affect the quality and distribution of the output data samples. Also, it will help to study how these tools could be used in order to build large, scalable datasets of synthetic facial data. Future works include, a study regarding the amount of data and variation needed in order to train StyleGAN effectively and a study regarding the relationship between the original samples used for training the StyleGAN models and the generated samples will be performed as well as examining the relationship between the generated samples.

ACKNOWLEDGMENT

REFERENCES

[1] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv Prepr. arXiv1703.10717*, 2017.

[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv Prepr. arXiv1511.06434*, 2015.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[5] S. Bazrafkan and P. Corcoran, "Versatile Auxiliary Classifier with Generative Adversarial Network (VAC+ GAN), Multi

[6] S. Bazrafkan and P. Corcoran, "Versatile Auxiliary Regressor with Generative Adversarial network (VAR+ GAN)," *arXiv Prepr. arXiv1805.10864*, 2018.

[7] "Making Anime Faces With StyleGAN," 2019. [Online]. Available: https://www.gwern.net/Faces. [Accessed: 29-Jan-2020].

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv Prepr. arXiv1710.10196*, 2017.

[9] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," 2008.

[10] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv Prepr. arXiv1411.7923*, 2014.

[11] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, "Recent advances in face presentation attack detection," in *Advances in Computer Vision and Pattern Recognition*, 2019.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[13] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[14] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*, 2016, pp. 87–102.

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition.," in *bmvc*, 2015, vol. 1, no. 3, p. 6.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67–74.

[17] L. Wolf, T. Hassner, and I. Maoz, *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.

[18] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for cnn-based face verification," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2545–2554.

[19] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 464–473.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2746–2754.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[23] 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 *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[25] M. Marchesi, "Megapixel size image creation using generative adversarial networks," *arXiv Prepr. arXiv1706.00082*, 2017.

[26] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10215–10224.

[27] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv Prepr. arXiv1809.11096*, 2018.

Appendix E

Dataset Cleaning - A Cross Validation Methodology for Large Facial Datasets using Face Recognition

Dataset Cleaning - A Cross Validation Methodology for Large Facial Datasets using Face Recognition

Viktor Varkarakis
School of Engineeing
National University of Ireland Galway
Galway, Ireland
v.varkarakis1@nuigalway.ie

Peter Corcoran
School of Engineering
National University of Ireland Galway
Galway, Ireland
peter.corcoran@nuigalway.ie

Abstract— In recent years, large "in the wild" face datasets have been released in an attempt to facilitate progress in tasks such as face detection, face recognition, and other tasks. Most of these datasets are acquired from webpages with automatic procedures. As a consequence, noisy data are often found. Furthermore, in these large face datasets, the annotation of identities is important as they are used for training face recognition algorithms. But due to the automatic way of gathering these datasets and due to their large size, many identities folder contain mislabeled samples which deteriorates the quality of the datasets. In this work, it is presented a semi-automatic method for cleaning the noisy large face datasets with the use of face recognition. This methodology is applied to clean the CelebA dataset and show its effectiveness. Furthermore, the list with the mislabelled samples in the CelebA dataset is made available.

Keywords— face datasets, mislabeled identities, noisy samples, clean face dataset, semi-automatic cleaning, CelebA

I. INTRODUCTION

In the last few years, Convolutional Neural Networks (CNNs) have significantly enhanced the performance of the state-of-the-art methods in many areas, including face recognition [1]–[3]. New CNN architectures are released frequently along with new learning methodologies that push the limits of face recognition [4]–[6]. But this success is also due to the recent Big Data era that has emerged, which allows creating large face datasets with real images harvested from the Internet [7][8]. Generally, large face datasets are built in a semi-supervised way using image-search engines and thus prone to bad data samples due to mislabelling and poor image quality of some samples [9]. Often, the number of these bad data samples is not statistically significant for a particular task and they can be ignored, but, in other cases a small number of bad data samples can become quite significant and lead to poor training outcomes [10]. Therefore, cleaning the datasets from mislabelled samples is desirable for some use cases.

Consider for example a large dataset which is used to train a facial image generator, e.g. StyleGAN[11]. If the training dataset contains some mislabelled identities – i.e. wrong identity is assigned to a person – this is not critical for training a GAN that can create realistic faces as these mislabelled data samples still represent 'good' samples of facial images. However, if the task at hand switches to training a CNN to perform facial recognition, distinguishing between multiple identities, these mis-labelled samples are now 'bad' and if there are sufficient such data samples, the performance of the resulting face recognition CNN will be sub-optimal [10].

However, identifying mislabelled facial images automatically without human supervision is a very challenging task. This is due to the extreme variations of the facial images captured in the wild which can result in mis-labelling one's identity [9]. Also, it has been shown that large face datasets can typically have a noise ratio of bad data samples higher than 30% [12].

In few works, procedures and ways are described in order to minimize bad data samples when creating the dataset [10], [12]. Other researchers have tried in various ways to clean the noisy data samples from such large datasets. In [13], an anchor face that had the most neighbours was selected and a maximal subgraph starting from this anchor was regarded as the cleaning result. The authors in [14], proposed a three-stage graph-based method to clean the large face datasets using a community detection algorithm. Although these methods can clean a large part of the datasets, they have some limitations. After the cleaning procedure, the datasets may, either lack diversity as many variations are treated as outliers or the size of the dataset has decreased quite significantly due to the rigorous constraints imposed by the cleaning process.

Finally, some researchers have employed manual annotators, and have succeeded in constructing a variety of face datasets where most images are correctly labelled such as [15][16], but this approach requires significant human effort with overlapping of the data annotation to achieve a consensus on more difficult samples. It also remains prone to human error and variations in human judgement, especially on 'difficult' samples.

This work introduces a semi-automatic methodology to find and remove mislabelled samples. from large facial datasets. Such facial datasets have many practical applications in building state-of-art multimedia experiences. A methodology for improving the quality of facial data samples in such datasets is an important tool for multimedia system & content developers.

This method described here utilizes a state-of-the-art face recognition (FR) model in order to detect the outliers within of a facial dataset which is organized with multiple classes of facial identity. Based on the intra-class comparisons of the samples, the images that produce low-confidence result are considered as outliers and examined manually. This is caused by either mislabelled samples or the samples which are difficult intra-class images for the FR model. This method does not dramatically reduce the size of the original dataset or reduces the diversity of the dataset. This method has been tested on a large facial dataset and the results are presented in this paper.

In the next sections the related literature is presented, followed by a description of the cleaning methodology for

facial identity datasets. Some examples of mislabelled data samples are described, and we have identified some common labelling errors across all of the dataset we have processed to date. Finally, results arising from an application of the methodology to the full CelebA dataset are presented and discussed and the list with the mislabelled samples are given in [1].

II. RELATED LITERATURE

In the following section an overview of publicly available facial datasets used for face recognition purposes are presented. Furthermore, as a face recognition model (FR) is used in the methodology, some the state-of-the-art face recognition algorithms are described shortly.

A. Publicly Available Facial Datasets

1) CASIA-WebFace

CASIA-WebFace [17] is one of the first large public facial datasets. It contains 10.575 identities with a total of 494.414 samples. The identities belong to celebrities and are collected from the IMDb website. The size of the dataset makes it suitable for training on the face recognition task and is frequently used throughout the literature.

2) CelebFaces

The CelebFaces+ dataset [18] 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. The dataset might also be known as CelebFaces Attributes Dataset (CelebA) [19] where the samples form CelebFaces+ are annotated with 5 landmark locations and for 40 binary attributes, providing valuable information for the researchers

3) VGGFace & VGGFace2

The VGG datasets are released from the Visual Geometry Group from the University of Oxford. The VGGFace [2] dataset was released in 2015 and contains 2.6M samples from 2,622 people. VGGFace was released similarly to CASIA-WebFace mainly for training purposes. In 2018, the VGGFace2 [20] was released which consists of 3.31M samples from 9,131 celebrities. The images were downloaded from Google Image Search. The image samples from VGGFace2 cover a wider range of different ethnicities, professions and age compared to VGGFace. Furthermore, all the samples have been captured "in the wild" thus giving the dataset a desirable variation with respect to pose, lighting and occlusion conditions as well as emotions. The dataset can be used for training and testing purposes as it is divided to a train and test set. Finally, VGGFace2 provides annotations regarding the pose and the age of its samples which can be useful for researchers.

4) Ms-Celeb-1M

The Ms-Celeb-1M dataset [9] was created and published in 2016 by Microsoft. It is the largest publicly available face recognition dataset with over 10M samples from 100K identities. The dataset is suitable for training and testing purposes.

B. Face Recognition Algorithms

Below a few state-of-the-art CNN based face recognition algorithms, are introduced.

1) DeepFace

In 2014, Facebook published DeepFace [3]. DeepFace at the time achieved state-of-the-art accuracy (97.35%) on the famous LFW benchmark. DeepFace introduced a new alignment, employing explicit 3D face modelling in order to apply a piecewise affine transformation. Furthermore, to achieve such a performance they trained a nine-layer deep neural network with their in-house datasets which consists of 4 million face samples from more than 4,000 identities.

2) FaceNet

In 2015, Google introduced FaceNet [1] and achieved accuracy of 99.63% on the LFW benchmark. FaceNet was trained on 200M images from 8 million subject. Furthermore, they introduced the triplet loss function. It requires the face triplets (an anchor, a sample of the same class as the anchor and a negative sample), and then it minimizes the distance between an anchor and a positive sample of the same identity and maximizes the distance between the anchor and a negative sample of a different identity.

3) ArcFace

ArcFace [5] was published in 2018. It pushed the limits of the LFW benchmark even further as it achieved 99.83% accuracy. It also achieved state-of-the-art results on the MegaFace Challenge. Finally, the authors proposed a new loss function, additive angular margin, to learn highly discriminative features for robust face recognition

III. METHODOLOGY

The following section describes, the methodology for finding and removing mislabelled samples in identity folders from face dataset. This methodology comprises three main stages. Initially a FR model is utilized to get an embedding from all the images. After, using the embeddings, a score for all the positive pairs from the face dataset is calculated. In the second stage the worst 2%-3% of identities of the dataset is thresholded as outliers. Finally, through a selection method, possible mislabelled samples from the thresholded identities are selected to be manually examined. The resulting data sample pairs – typically not more than a few thousand even on a large dataset - are manually examined.

A. Scores from Face Recognition (FR) Model

Firstly, the FR model is trained (or fine-tuned) on the original dataset which is to be cleaned. Note that following an initial cleaning of the dataset, the FR can be further fine-tuned by retraining on the cleaned dataset. Thus, several iterations can be run to further improve the cleaning. This methodology leverages the power of the FR model to distinguish samples of different facial identities. The FR model must have a very good performance on the examined dataset in order to be able to detect outliers efficiently. If the FR model does not have high performance on the dataset, correctly labelled samples will be easily considered as outliers. Training / fine-tuning the FR on the dataset that will be examined, has a trade-off, as there is the possibility that the FR model will learn to classify a sample "correctly" even if it is mislabeled. Although it is assumed that the mislabeled entities comprise a very small percentage of the database and does not have a

[1] https://github.com/C3Imaging/Deep-Learning-Techniques/tree/clean-celebA

big effect on the final model. This gives the FR model the opportunity to learn the most representative embeddings during training and mislabeled samples will be treated as outliers.

It is not a necessary step to train / fine-tune the FR on the examined dataset as FR can perform well on a dataset even if it has been trained on a different one. Although this is recommended as the FR model will thus be better optimized for the dataset that is selected to be cleaned.

Next, the embeddings for all the images are produced from the FR model. For all the positive pairs for each identity, the score is calculated using the embeddings (*pair score*). The selection of the score depends on the way the FR model was optimized (Euclidean distance, cosine similarity etc.), as models can be optimized with different losses. The proposed methodology utilizes the Euclidean distance to measure the difference between the embeddings of two images.

For each identity the scores from all the possible positive pairs are calculated and the worst score is selected as the score of the identity *(id score)*. In this way we take into consideration all the intra-class samples and it enables us to examine how good are the embeddings of the FR model produced for each identity.

B. Outlier Selection

After the procedure described above, each identity is assigned with an *id score*. The 2-3% of the identities with the worst *id score*, are thresholded and marked as outliers.

It is chosen to examine only the top 2-3% of the dataset as we do not want to dramatically reduce the size of the original dataset or reduces its diversity. It is desired to only to remove the most obvious outliers. There is a high possibility that the mislabelled samples are discriminated after thresholding since the FR model was not able to produce embeddings that are close enough. Although, that does not necessary means that all the samples from these identities are mislabeled.

In order to fine-grain the selection of the possible mislabeled samples from the thresholded identities, the images from the pairs that produced a low-confidence *pair score* are targeted. To do that, another threshold is defined (*pair threshold*) which is selected, based on the average value of all the *id scores* from the identities. If a pair has produced a *pair score* worse than the *pair threshold,* then the images of the pair are recorded. Also, it is noted how often an image participated in pairs that produced *a pair score,* worse than the *pair threshold*. Therefore, in this step the samples with the biggest internal embedding distance are selected as mislabeled.

To summarize, two thresholding procedures are being implemented at this step. The first one is being implemented to threshold the identities that might have mislabeled samples in their folder. The second thresholding is implemented in order to fine-grain which samples from the thresholded identities might be the mislabeled ones.

C. Selection of Samples for Visual Examination

As, mentioned in the previous section, the identities that might contain mislabeled samples followed by the image pairs are thresholded. Also, *image frequency* was introduced as the number of times that an image participated in a pair and had a *pair score* more than the *pair threshold.*

Based on the *image frequency* of a sample, it is determined whether it will be manually examined or not. For each identity the samples that are manually examined are selected using the following procedure. The number of pairs that have *pair score* worse than the *pair threshold* is calculated (*NoP*). Then the samples are sorted in a descending order based on their *image frequency*. Starting from top to bottom a sample is selected. Every time a sample is selected, its *image frequency* is subtracted from *NoP*. Samples are selected till *NoP* is equal or less to 0. Finally, these samples are manually examined in order to identify the mislabeled ones.

The initial experiments using the proposed methodology indicated that there are 3 common types of mislabeling in the identity folders:
a) One main identity with 1 to n, mislabeled samples in the folder
b) An identity folder with n mislabeled samples and without one of the different identities having a stronger presence than the others. By stronger presence, it is meant to have enough samples to create an identity folder (more than 4-5 samples).
c) Two identities in the same folder.

IV. EXPERIMENTS ON CELEBA

In the next section, the methodology described is applied to clean the CelebA dataset from mislabeled samples and the result are presented with examples from each mislabeling type.

A. Scores from FR model on CelebA

The FR model selected for this set of experiments can be found here [2]. This is an unofficial TensorFlow implementation of FaceNet [1], built on ideas from [2]. This FR model/ implementation was selected for two main reasons. The first, being its availability and its ease of use. The second is the fact that it provides a pretrained model which reports state-of-the-art performance in the LFW test set [7]. For the purpose of this research, the available pre-trained model is fine-tuned on the CelebA dataset. The FR model's architecture is an Inception ResNetv1 [21]. The employed pretrained model that is trained with SoftMax, on the VGGFace2 dataset [20]. The input size of the network is an $160x160$ image and the output is a 512-embeding.

The reason for the fine-tuning is for the FR to be more dataset specific and have a higher performance on the dataset that will be examined. In the fine-tuning process, the same configurations as in training were used, with a reduced learning rate. For more information regarding the training of the FR model and data preparation see [2].

In Fig.1 the ROCs on the CelebA dataset is presented for the models before and after fine-tuning. The ROC curves, shows that the FR model after fine-tuning on the CelebA,

---

[2] https://github.com/davidsandberg/facenet

performs better than the pretrained model. Therefore, it will point to the identity folders that may have mislabeled samples more effectively. This is because the performance is increased, resulting in a lower false positive error. This also illustrates the need for the FR model to be trained / fine-tuned on the examined dataset.

After fine-tuning is completed, the 512-embedding for all the samples of CelebA dataset are calculated. The score used for this set of experiments is the Euclidean distance between the embeddings. For each identity, the scores from all the possible positive pairs are calculated and the worst score (in this case the highest Euclidean distance) is selected as the score of the identity *(id score)*.
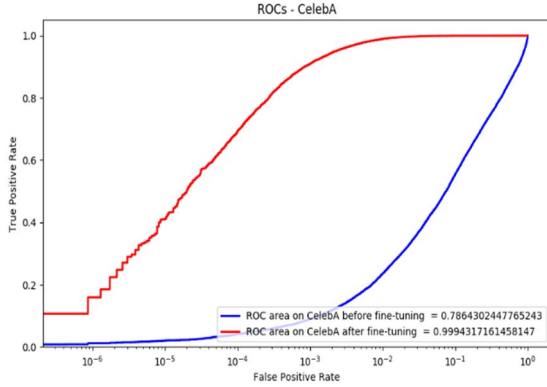

Fig. 1. The ROC curves on the CelebA dataset. The blue line, shows the performance for FR model before fine-tuning and the red for the FR model after-finetuning.

B. Identifying possible mislabeled identity folders

As mentioned in the section B of the Methodology, in two thresholding procedures take place at this step.

In the first thresholding procedure, the identities with the worst 3% *id score* (in this case with highest Euclidean distance) are thresholded. These identities are considered as outliers, as they might contain mislabeled samples in their folders. This means 310 identities.

Afterwards, a second thresholding takes place. This is implemented in order to fine-grain the selection of the possible mislabeled samples that may exist in the thresholded identity folders (section III-B).

First, the *pair threshold* is calculated as described in the Methodology (section B) which is the average of all the *id scores* from all the identities and is equal to 1.

Therefore, for the thresholded identities, all the positive image pairs that have Euclidean distance more than 1 (*pair threshold*), are recorded. Also it is noted how often an image exist in pair with *pair score,* worse than the *pair threshold,* and is defined as *image frequency*. In Fig.2 illustrates the distribution of the *id score* from all the identities.

C. Results

After the two thresholding operations (initially for the identities and afterwards for its samples), the *image frequency* is finally calculated. As mention in the section C, of the Methodology, *the image frequency* is used to determine
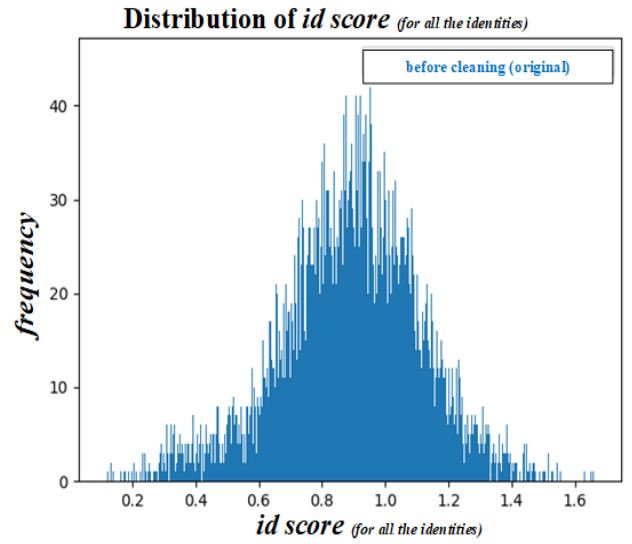

Fig. 2. The distribution of the *id score* from all the identities from the CelebA dataset before applying the method proposed for cleaning the dataset.

which samples will be manually examined for being possible mislabeling.

Applying the methodology, the three types of mislabeling in the identity folders appeared as well as cases where the methodology flagged an identity folder which by examining its samples, it did not have any mislabeling, but it contained samples with high variation. Below some examples are presented for each case, along with the different actions that were chosen for cleaning the dataset, depending on the mislabeling type.

1. Two identities in one folder

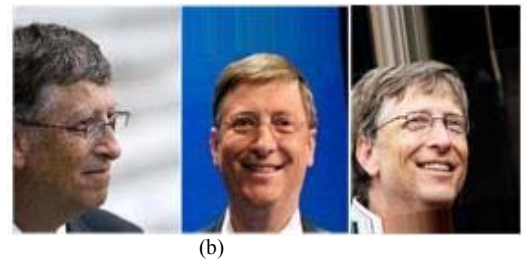Fig. 3. Example of a mislabeled folder in CelebA, which contains two different identities.

In this case the mislabeling type of having two identities in the same folder was detected using the described technique. Fig. 3a and Fig. 3b are samples, obviously from 2 different identities but found on the same identity folder. In this cases, one identity is retained, and the samples of the other identity are removed

2. One identity with n mislabeled samples

In this case the technique used, flagged an identity folder which contained samples from one identity but also some mislabeled samples. Fig. 4a, shows the mislabeled samples existing in the identity folder. Fig. 4b shows the samples belonging to the same identity.

In these cases, the mislabeled samples were deleted. In case that the mislabeled samples were many and the main identity was left with less than 2-3 samples, the folder was as well removed.

(a)

(b)

Fig. 4. Example of an folder in CelebA, which contains one main identity with 1 to n, mislabeled samples in the folder.

3. An identity folder with n mislabeled samples

Fig. 5. Example of an folder in CelebA, with n mislabeled samples and without one of the different identities having a stronger presence than the others.

In this example a folder with n different identities were detected, as it can be seen in Fig.5. In this type of mislabeling, where a folder had n mislabeled samples without one identity having a stronger present (By stronger presence, it is meant to have enough samples to create an identity folder (more than 4-5 samples)), the whole identity folder was deleted.

4. High variation in an identity folder

Finally, there were cases were the methodology indicated an identity folder as an outlier. Though by examining its samples, it was detected that the samples belong to the same identity. The technique indicated this identity folder as an outlier due to its high variation. In Fig. 6, such an example is illustrated. In case a folder was considered an outlier without having any mislabeled samples, no further action was taken, and its samples were retained.

Fig. 6. Example of an folder in CelebA, shows a case where the methodology was unsuccessful. As it identified this folder as outlier with possible mislabeled samples but the folder just contained difficult intra-class samples due to variation for the FR model.
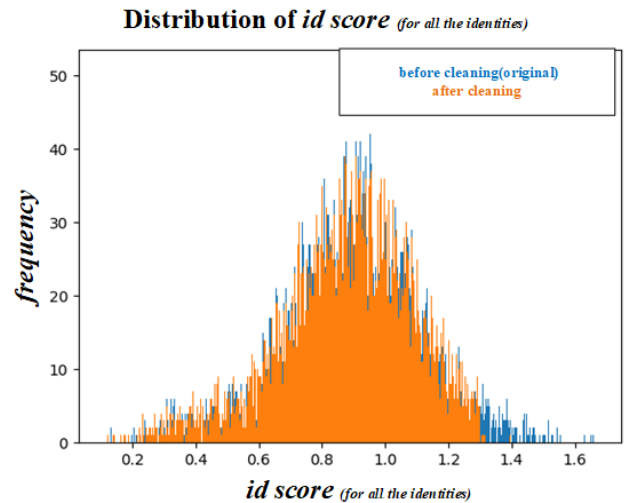

Fig. 7. The distribution of the *id score* from all the identities from the CelebA dataset before (blue) applying the method proposed for cleaning the dataset and after (orange). The distribution of the *id scores* was reduced when the mislabeled samples were removed, showing the effect of cleaning in the dataset with the proposed method.

In total, from the 310 identities that were flagged using the proposed method only 9 of them did not have any mislabeling samples. The other 301 identities selected contained mislabeled samples. In Fig. 2, the distribution of the *id score* for all identities is shown before cleaning the CelebA dataset. In Fig. 7 the distribution of the *id score* is shown after the cleaning of the CelebA dataset, in comparison with the initial distribution of the *id score* from Fig2. It can be seen from the Fig.7, that majority of the high scores were

due to the mislabeled samples, as after removing the mislabeled samples the distribution of the *id score* was reduced.

The CelebA dataset as mentioned earlier, consist of 202,599 images from 10,177 identities. After applying the methodology for finding and removing mislabeled samples as described earlier, it remains with 197,477 samples from 9,996 identities. The list with the mislabeled samples is publicly available in [1].

V. Discussion and Future Work

In the sections above, a (semi-automatic) technique for identifying and removing mislabeled samples in terms of identity is described. The technique utilizes a face recognition model trained / fine-tuned on the examined dataset in order to discover outliers in an identity folder that shall be examined as it is possible to contain mislabeled face samples. This methodology was applied to clean the CelebA dataset and the results are presented in section IV-C. In addition, the list with the mislabeled samples can be found in [1]. This technique can be applied to any face dataset annotated with identities in order to "clean" it so that the dataset can be used with more certainty as a considerable number of mislabeled samples will be eliminated.

In this preliminary work our main goal has been to demonstrate the effectiveness of the methodology to provide a minimal curation of the dataset. In other words, we seek to retain as many of the original data samples as possible to ensure that the diversity of the original dataset is preserved. There is still a lot of work to apply these techniques across additional large datasets and to further automate the methodology and develop additional analysis tools and quality metrics to fully demonstrate its capability to improve the quality of these datasets.

Also, this technique will be examined in order to observe the influence and how to achieve the best configuration for setting the identity and pair threshold, as the tuning of this threshold has not been explored in detail in this preliminary work. Furthermore, this methodology will be used to identify mislabeled samples in other face datasets. The methodology should also be compared with some datasets that have been manually cleaned and we are currently signing some license agreements to gain access to a number of such 'clean' datasets. It is expected that some comparisons can be provided for presentation at QoMEx 2020.

Finally, it would be useful to automate additional aspects of the cleaning process and approaches to reduce the computational complexity of the methodology. A number of these will be explored, working in collaboration with other researchers later this year.

Acknowledgment

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition.," in *bmvc*, 2015, vol. 1, no. 3, p. 6.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[4] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[7] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th international conference on computer vision*, 2009, pp. 365–372.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*, 2016, pp. 87–102.

[10] I. Gallo, S. Nawaz, A. Calefati, and G. Piccoli, "A Pipeline to Improve Face Recognition Datasets and Applications," in *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2018, pp. 1–6.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[12] F. Wang *et al.*, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[13] S. Ding, J. Wu, W. Xu, and H. Chao, "Automatically building face datasets of new domains from weakly labeled data with pretrained models," *arXiv Prepr. arXiv1611.08107*, 2016.

[14] C. Jin, R. Jin, K. Chen, and Y. Dou, "A community detection approach to cleaning extremely large face database," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[15] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 464–473.

[16] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in face detection and facial image analysis*, Springer, 2016, pp. 189–248.

[17] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv Prepr. arXiv1411.7923*, 2014.

[18] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, "Recent advances in face presentation attack detection," in *Advances in Computer Vision and Pattern Recognition*, 2019.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67–74.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

Appendix F

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

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

VIKTOR VARKARAKIS[ID][1], **(Graduate Student Member, IEEE),**
SHABAB BAZRAFKAN[2], **(Student Member, IEEE), GABRIEL COSTACHE**[3],
AND PETER CORCORAN[ID][1], **(Fellow, IEEE)**
[1]Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
[2]Imec Vision Laboratory, Department of Physics, University of Antwerp, 2000 Antwerp, Belgium
[3]Xperi, Galway, H91 V0TX Ireland

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

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

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu[ID].

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 quantitively 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 filed. 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.[1] 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[2] 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 the 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

[1] https://github.com/deepinsight/insightface
[2] https://github.com/davidsandberg/facenet

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).
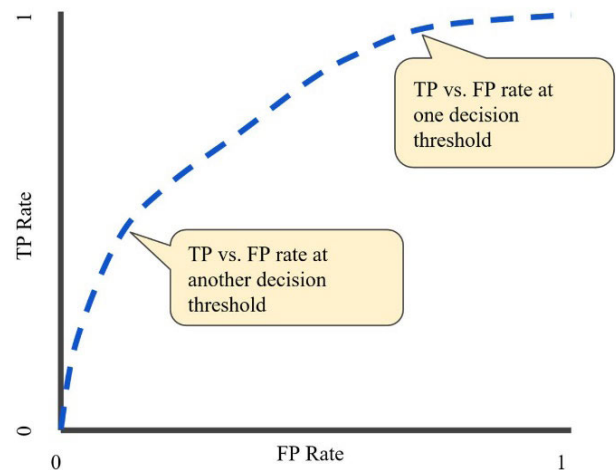

FIGURE 1. A typical ROC curve, TP vs. FP rate at different classification thresholds [53], [54].

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.
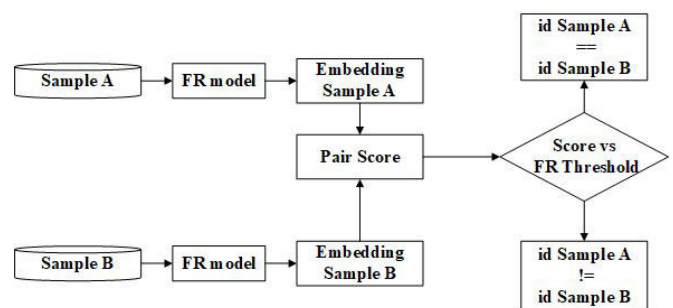

FIGURE 2. The workflow of Thresholding Technique.

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.
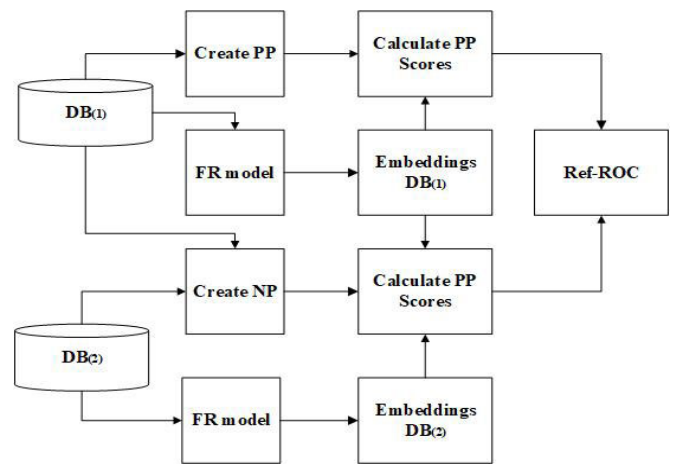

FIGURE 3. The workflow of computing a *Ref-ROC*.

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
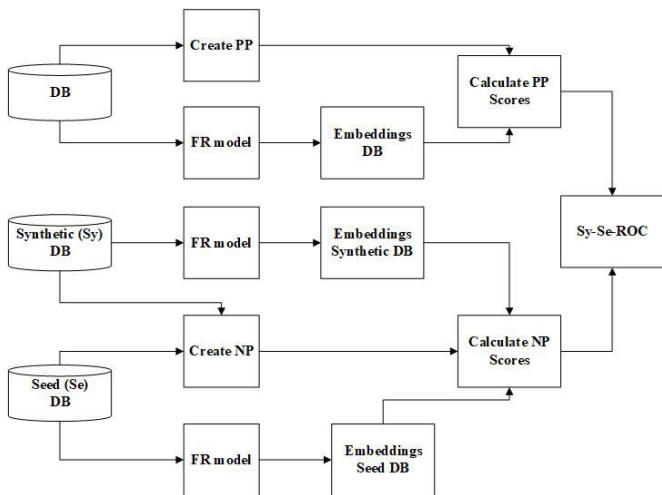
FIGURE 4. The workflow of computing a Sy-Se-ROC.

FIGURE 5. The workflow of computing a Sy-ROC.

NPs 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 and it is concluded that the generated data samples are not unique when compared with the seed dataset in terms of identity.

3) IDENTITY UNIQUENESS AMONG THE SYNTHETIC DATA (SY VS SY) - ROC CURVES COMPARISON

To examine the identity uniqueness among the generated synthetic data, a ROC curve is computed using these samples (Sy-ROC) and compared against the Ref-ROC curve. The Sy-ROC curve uses the same PPs as the Ref-ROC, but the NPs consist of synthetic data pairs. The scores of the PPs and NPs are calculated and used to compute the Sy-ROC. The workflow of computing a Sy-ROC is given in Fig 5.

When the Sy-ROC is compared against the Ref-ROC, the ROC curves differ only in the NPs which in this case are synthetic data pairs. The synthetic data pairs are treated as NPs as they don't have an identity label. Similarly, as in section IV-B-2, when the Sy-ROC is at similar or higher
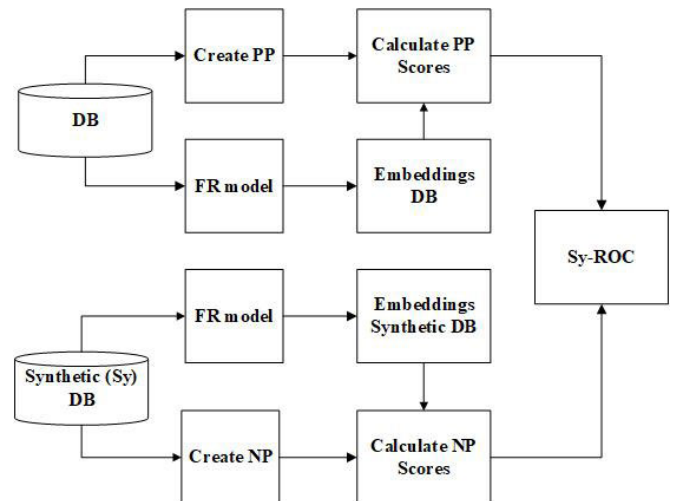
levels than the Ref-ROCs, it can be concluded that the identity uniqueness of the synthetic pairs is similar or higher as the one of real samples from different identities. Thus, showing that the generated synthetic data are unique in terms of identity, which is desirable. Conversely, if the Sy-ROC lies below the Ref-ROC, then the synthetic data pairs show a lower statistical identity uniqueness between them compared to that of real samples, concluding that the generated synthetic data are not unique in terms of identity.

C. THRESHOLDING 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.
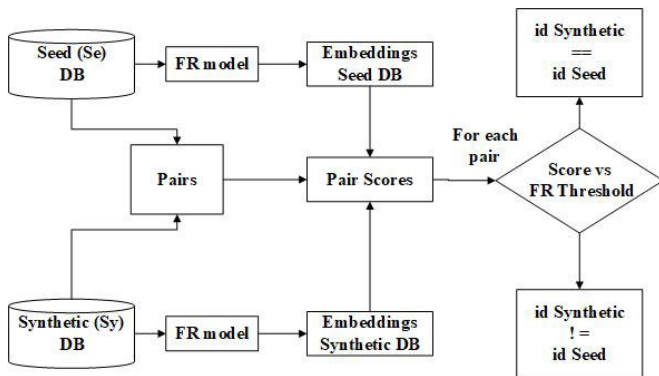

FIGURE 6. The workflow of the Thresholding Technique to examine the identity uniqueness between synthetic and seed samples (Sy vs Se).

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
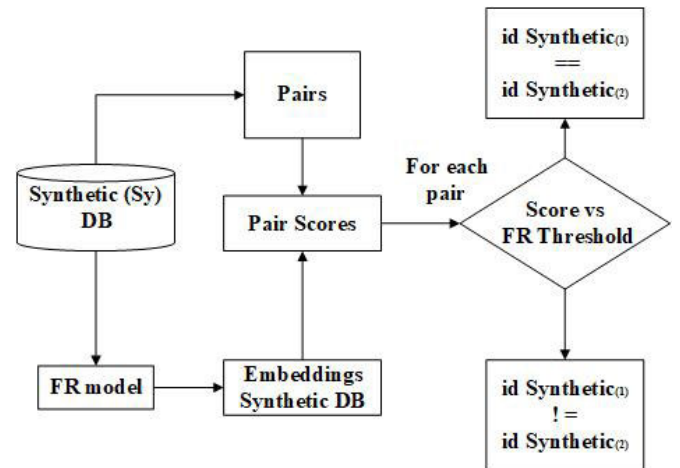

FIGURE 7. The workflow of the Thresholding Technique to examine the identity uniqueness among the synthetic samples (Sy vs Sy).

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:

$$REFP = \left(\frac{NoP}{NoC} \right) / FPR \qquad (1)$$

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.
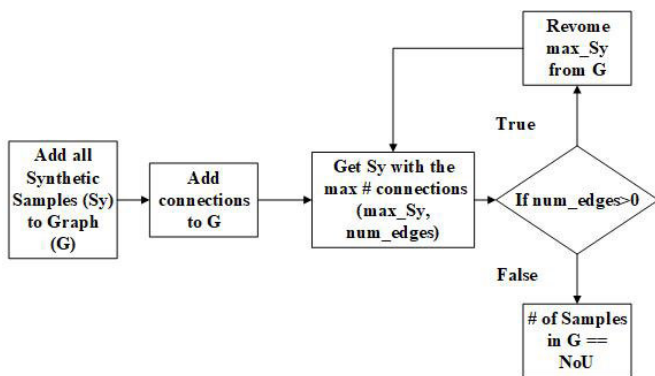

FIGURE 8. The workflow of calculating the NoU within a synthetic set of samples (Sy vs Sy).

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 Style-GAN [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.[3] 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''.

TABLE 1. The synthetic datasets created along with the truncation psi value used and information of their seed dataset.

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-0.1	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

---

[3]https://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.[1] The MTCNN [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.[4]

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.

TABLE 2. Datasets used to compute the pp and the np for each Ref-ROC curve as shown in Fig.3.

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

TABLE 3. The synthetic and the seed datasets used to computed each sy-se-roc curve, as shown in fig.4.

Name of Sy-Se-ROC	PP Dataset	NP Datasets	
		Seed DB	Synthetic DB
CelebA-0.5	CelebA	CelebA	Sy-CelebA-0.5
CelebA-0.7	CelebA	CelebA	Sy-CelebA-0.7
CelebA-1	CelebA	CelebA	Sy-CelebA-1.0
CelebA-HQ-0.5	CelebA	CelebA-HQ	Sy-CelebA-HQ-0.5
CelebA-HQ-0.7	CelebA	CelebA-HQ	Sy-CelebA-HQ-0.7
CelebA-HQ-1	CelebA	CelebA-HQ	Sy-CelebA-HQ-1
FFHQ-0.5	CelebA	FFHQ	Sy-FFHQ-0.5
FFHQ-0.7	CelebA	FFHQ	Sy-FFHQ-0.7
FFHQ-1	CelebA	FFHQ	Sy-FFHQ-1

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. 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.
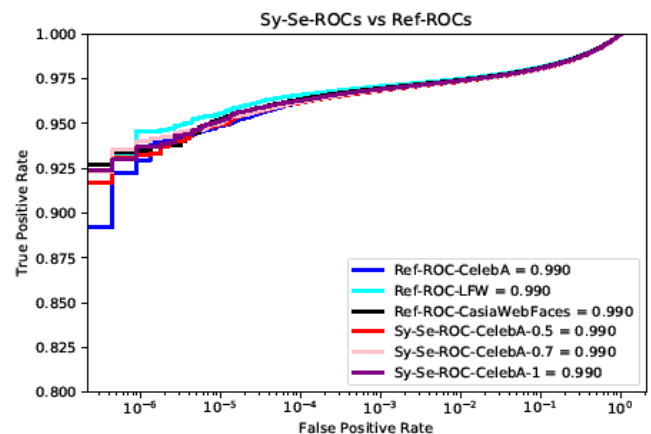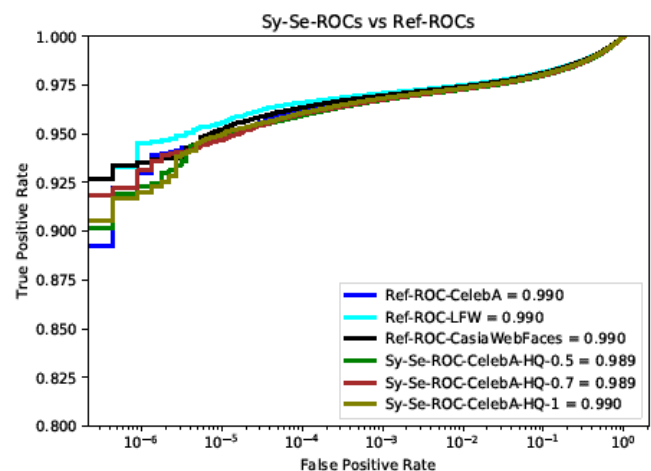

FIGURE 10. The Sy-Se-ROCs from StyleGAN-CelebA-HQ (Table 3) compared against the Ref-ROCs to determine the identity uniqueness between the synthetic and the seed samples.

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
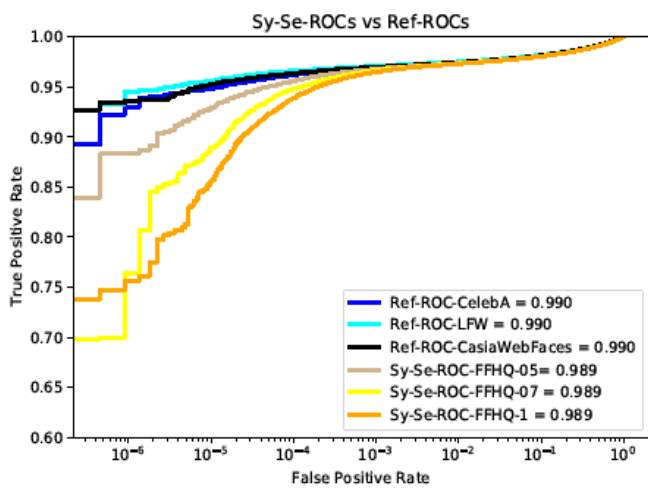
as in Fig.11 are presented but with the manual removal of several data pairs that consist of infants or young children. It is clear that these remain at similar levels or above the Ref-ROC and the behavior is now similar to Fig.9 and 10. This indicates that the identity uniqueness between the synthetic 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.

FIGURE 11. The Sy-Se-ROCs from StyleGAN-FFHQ (Table 3) compared against the Ref-ROCs to determine the identity uniqueness between the synthetic and the seed samples.
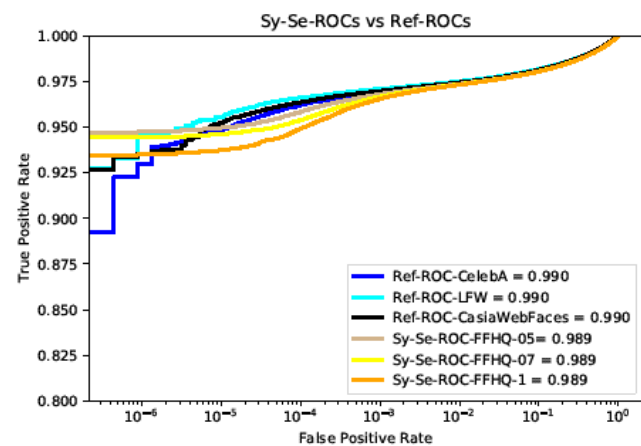

FIGURE 13. The Sy-Se-ROCs (Fig. 9,10 and 12) are compared between to examine which StyleGAN model performs better in generating synthetic samples that are unique when compared to their seed data.

FIGURE 12. The Sy-Se-ROCs from StyleGAN-FFHQ (Table 3), with pairs of infants/kids removed, are compared against the Ref-ROC.

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
		Synthetic DB
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
CelebA-HQ-1	CelebA	Sy-CelebA-HQ-1
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
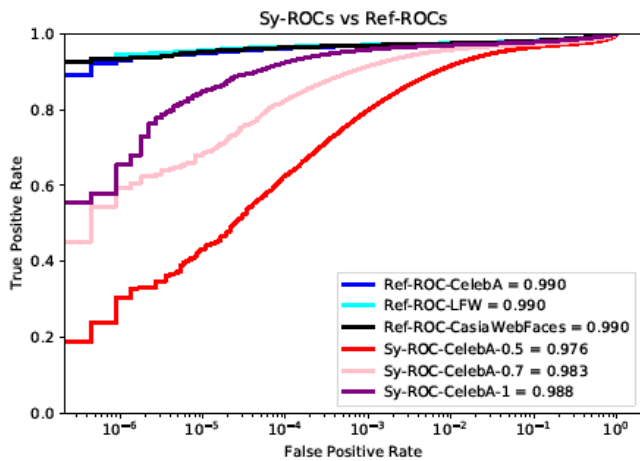
FIGURE 14. The Sy-ROCs from StyleGAN-CelebA (Table 4) compared against the Ref-ROCs to determine the identity uniqueness among the synthetic samples.
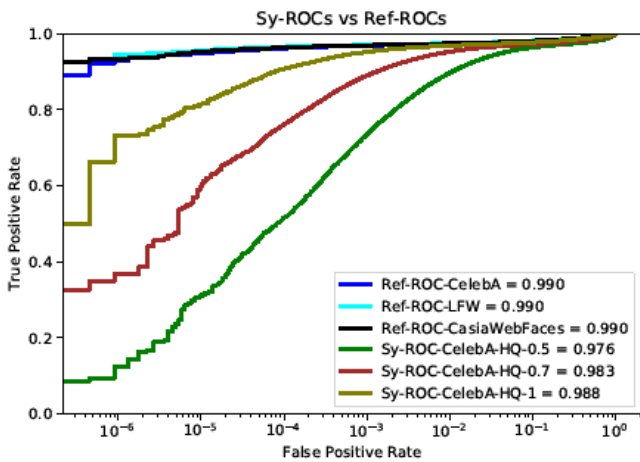
FIGURE 15. The Sy-ROCs from StyleGAN-CelebA-HQ (Table 4) compared against the Ref-ROCs to determine the identity uniqueness among the synthetic samples.
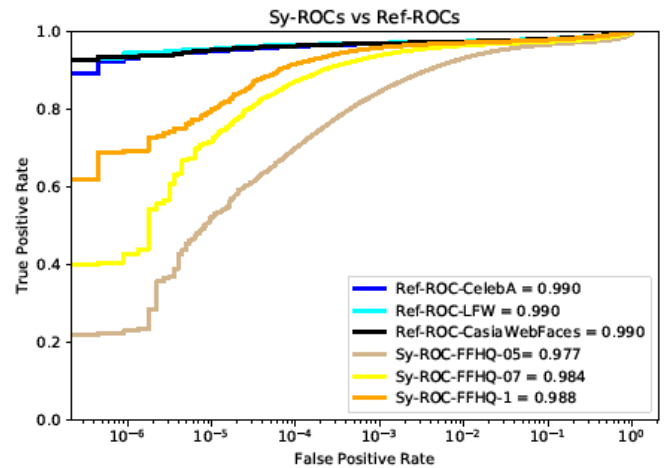
FIGURE 16. The Sy-ROCs from StyleGAN-FFHQ (Table 4) compared against the Ref-ROCs to determine the identity uniqueness among the synthetic samples.

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).
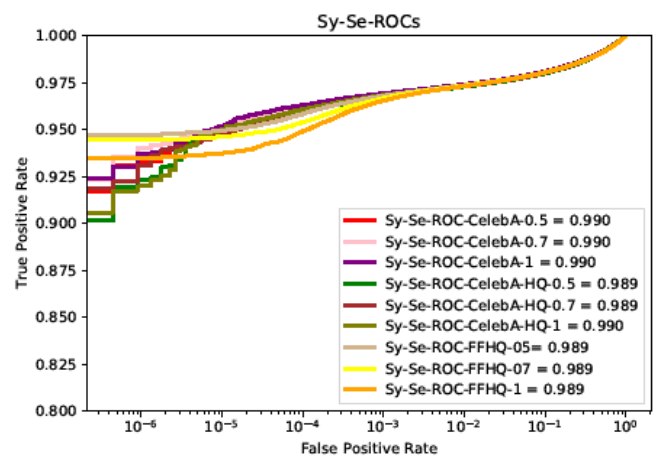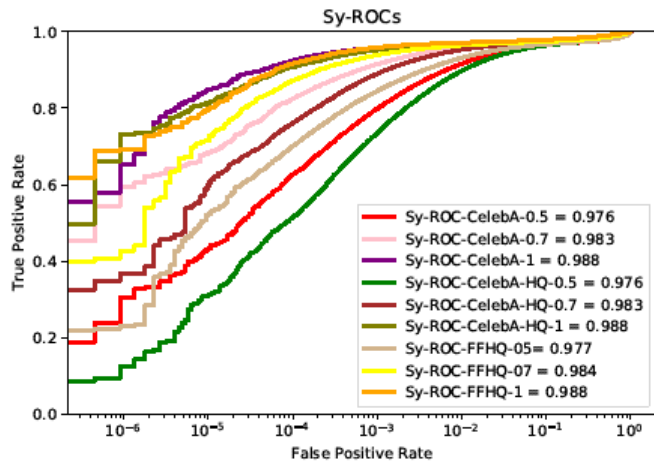
FIGURE 17. The Sy-ROCs (Table 4) are compared to examine which StyleGAN model performs better in generating synthetic samples with a unique identity.

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.

TABLE 5. The synthetics and seed dataset used to implement the Thresholding Technique for the Sy vs Se case.

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-CelebA-HQ-1	Sy-CelebA-HQ-1	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

TABLE 6. The REFP for each synthetic dataset for the sy vs se case.

Name of Comparison	FPR/ Threshold		
	4.46e-07/ 0. 4823	1.33e-06/ 0. 4280	1e-05/ 0.3815
Sy-Se-CelebA-0.5	0.05	0.47	0.68
Sy-Se-CelebA-0.7	0.05	0.47	0.64
Sy-Se-CelebA-1	0.07	0.46	0.56
Sy-Se-CelebA-HQ-0.5	0.13	1.02	1.31
Sy-Se-CelebA-HQ-0.7	0.21	1.01	1.30
Sy-Se-CelebA-HQ-1	0.40	1.23	1.16
Sy-Se-FFHQ-0.5	5.08	8.94	4.37
Sy-Se-FFHQ-0.7	23.52	30.30	11.51
Sy-Se- FFHQ-1	40.71	50.07	17.88

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 have a lower REFP than the ones generated with the StyleGAN-CelebA-HQ. This indicates that the StyleGAN-CelebA performs marginally better in generating samples that are unique from their seed dataset. No further examination is conducted regarding the synthetic datasets from the StyleGAN-FFHQ model as the samples of infants/young children pose challenges in computing the REFP correctly.

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), and 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.

TABLE 7. The synthetic datasets used to implement the Thresholding Technique for the Sy vs Sy case.

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

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.

TABLE 8. The REFP for each synthetic dataset for the Sy vs Sy case.

Name of Comparison	FRP/Threshold		
	4.46e-07/ 0. 4823	1.33e-06/ 0.4280	1e-05/ 0.3815
Sy-Sy-CelebA-0.5	12340.10	11298.56	3165.65
Sy-Sy-CelebA-0.7	1085.65	1277.62	447.18
Sy-Sy-CelebA-1	61.89	86.17	36.50
Sy-Sy-CelebA-HQ-0.5	20335.89	16165.28	4060.06
Sy-Sy-CelebA-HQ-0.7	2343.97	2410.03	749.60
Sy-Sy-CelebA-HQ-1	145.78	189.23	73.16
Sy-Sy-FFHQ-0.5	6521.42	6864.26	2122.12
Sy-Sy-FFHQ-0.7	365.85	469.84	182.83
Sy-Sy-FFHQ-1	128.31	129.44	42.37

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

TABLE 9. The number of samples with a unique identity (NoU) and the percentage (%) of them within each synthetic dataset.

Name of Comparison	FRP/Threshold		
	4.46e-07/ 0. 4823	1.33e-06/ 0.4280	1e-05/ 0.3815
Sy-Sy-CelebA-0.5	5691, (28.45%)	3286, (16.43%)	1935, (9.67%)
Sy-Sy-CelebA-0.7	12160, (60.8%)	8236, (41.18%)	5297, (26.48%)
Sy-Sy-CelebA-1	17835, (89.17%)	14876, (74.38%)	11117, (55.58%)
Sy-Sy-CelebA-HQ-0.5	4580, (22.90%)	2638, (13.19%)	1564, (7.82%)
Sy-Sy-CelebA-HQ-0.7	10091, (50.45%)	6498, (32.49%)	4053, (20.27%)
Sy-Sy-CelebA-HQ-1	16227, (81.13%)	12560, (62.8%)	8766, (43.83%)
Sy-Sy-FFHQ-0.5	5985, (29.92%)	3267, (16.33%)	1847, (9.23%)
Sy-Sy-FFHQ-0.7	14217, (71.08%)	9606, (48.03%)	6018, (30.09%)
Sy-Sy-FFHQ-1	18466, (92.33%)	16298, (81.49%)	12795, (63.97%)

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.[5]

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.

[5]https://github.com/C3Imaging/Deep-Learning-Techniques/tree/ Synthetic_Face_Datasets
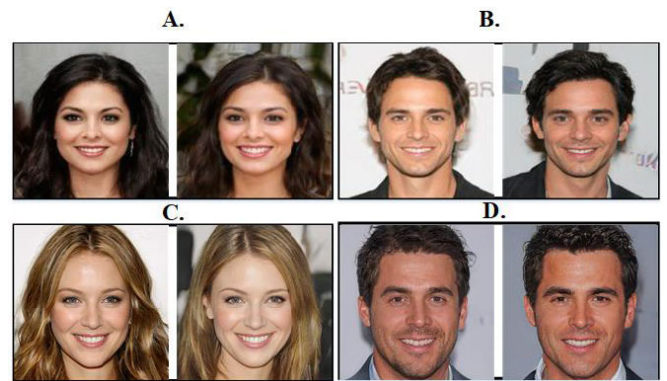
FIGURE 18. The synthetic pair samples have a similarity score above the FR threshold of this work (0.48). Thus, the synthetic samples have an identity similarity and therefore are not unique. The similarity score for the pairs is: A=0.72, B=0.70, C=0.75, D=0.74.

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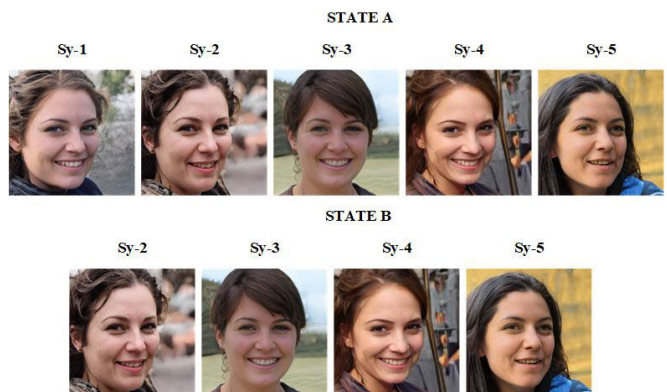
FIGURE 19. A cluster of synthetic samples which have an identity similarity between them (state A). The graph theory technique of section IV-C-5 removes the sample with the most identity connections resulting in four unique identities (state B). The corresponding similarity scores, and number of connections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

TABLE 10. The identity similarity score between the samples (Fig.19-state A) and the number of connections of each sample.

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

TABLE 11. The identity similarity score between the samples (Fig.19-state B) and the number of connections of each sample.

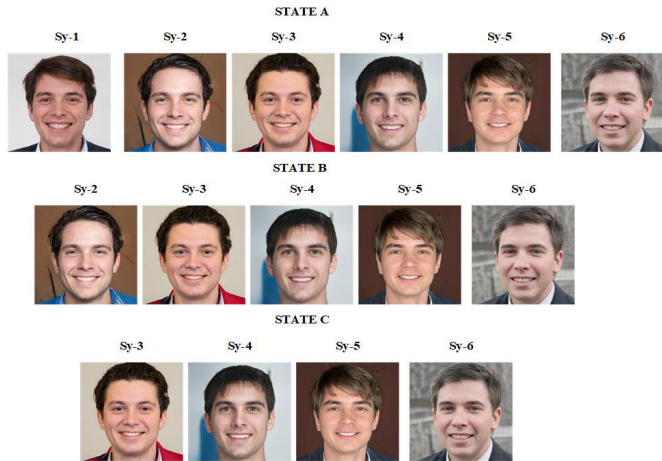| Samples	Sy-2	Sy-3	Sy-4	Sy-5	Con
Sy-2	-	0.39	0.30	0.36	0
Sy-3	0.39	-	0.28	0.23	0
Sy-4	0.30	0.28	-	0.29	0
Sy-5	0.36	0.23	0.29	-	0

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.

TABLE 12. The identity similarity score between the samples (Fig.20-state A) and the number of connections of each sample.

Samples	Sy-1	Sy-2	Sy-3	Sy-4	Sy-5	Sy-6	Con
Sy-1	-	**0.57**	**0.56**	**0.50**	**0.49**	**0.49**	**5**
Sy-2	**0.57**	-	**0.51**	**0.53**	0.37	0.38	**3**
Sy-3	**0.56**	**0.51**	-	0.36	0.38	0.40	2
Sy-4	**0.50**	**0.53**	0.36	-	0.35	0.32	2
Sy-5	**0.49**	0.37	0.38	0.35	-	0.24	1
Sy-6	**0.49**	0.38	0.40	0.32	0.24	-	1

TABLE 13. The identity similarity score between the samples (Fig.20-state B) and the number of connections of each sample.

Samples	Sy-2	Sy-3	Sy-4	Sy-5	Sy-6	Con
Sy-2	-	**0.51**	**0.53**	0.37	0.38	**2**
Sy-3	**0.51**	-	0.36	0.38	0.40	1
Sy-4	**0.53**	0.36	-	0.35	0.32	1
Sy-5	0.37	0.38	0.35	-	0.24	0
Sy-6	0.38	0.40	0.32	0.24	-	0

TABLE 14. The identity similarity score between the samples (Fig.20-state C) and the number of connections of each sample.

Samples	Sy-3	Sy-4	Sy-5	Sy-6	Con
Sy-3	-	0.36	0.38	0.40	0
Sy-4	0.36	-	0.35	0.32	0
Sy-5	0.38	0.35	-	0.24	0
Sy-6	0.40	0.32	0.24	-	0

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.

Algorithm 1 Graph Approach Which Allows to Quantify the Number of Samples With a Unique Identity (NoU) Within a Synthetic Dataset

Find the maximum number of SY with a unique identity(NoU) within a set of generated synthetic samples. The *name_of_nodes*, is a list and contains all the SY samples. The *Sy_edges* is a list and contains all the pairs of SY that have an identity similarity with e.g.: Sy_edges = [(SYx, SYy),....] show that the SYx and SYy have an identity similarity and therefore connection (edge) for the graph.

- Create the graph G:

G=Graph()

- Add all nodes to the graph G

G.add_nodes_from(name_of_nodes)

- Add all the edges (nodes that are connected to the graph G)

G.add_edges_from (Sy_edges)

- List the nodes (max_nodes) with the most connections and the value of the connections (num_edges)

max_nodes, num_edges= max(G.degree())

- Check if the max number of connections is more than 0 otherwise, remove the max_nodes from the graph G. (if len(max_nodes)>1 choose randomly a node a value and remove the corresponding graph from the graph)
- Repeat that till the num_edges <=0

While (num_edges>0):

    rand= random integer between 0 and len(max_nodes)

    G.remove_nodes_from ([max_nodes [rand]])

    max_nodes, num_edges= max(G.degree())

end while

- When over the maximum number of SY with a unique identity NoU) is the number of remaining nodes which don't have a connection with other nodes.

NoU= len(G.nodes())

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.
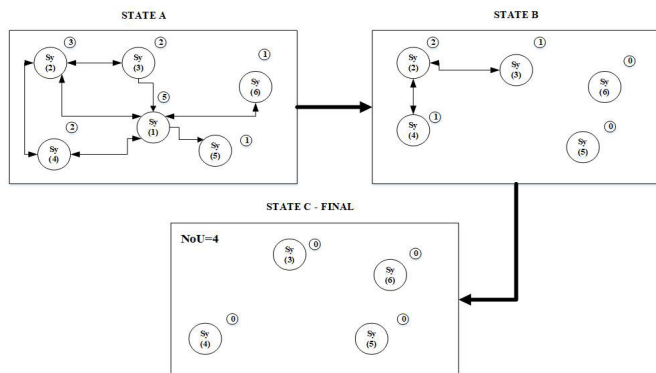

FIGURE 21. A graph containing several synthetic samples. The arrows show if two samples have an identity similarity between them and next to them is the number of connections of each sample. The Algorithm 1 is applied till the remaining samples do not have any connection between them which gives the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

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

Name of Comparison	FPR/ Threshold			NoC
	4.46e-07/ 0. 4823	1.33e-06/ 0. 4280	1e-05/ 0.3815	
Sy-Se-CelebA-0.5	90	2,533	27,572	4B
Sy-Se-CelebA-0.7	94	2,552	25,915	4B
Sy-Se-CelebA-1	133	2,459	22,771	4B
Sy-Se-CelebA-HQ-0.5	36	820	7,912	600M
Sy-Se-CelebA-HQ-0.7	57	807	7,837	600M
Sy-Se-CelebA-HQ-1	109	984	7,008	600M
Sy-Se-FFHQ-0.5	3,176	16,658	61,295	1.4B
Sy-Se-FFHQ-0.7	14,688	56,427	161,285	1.4B
Sy-Se-FFHQ-1	40.71	50.07	17.88	1.4B

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.

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			NoC
	4.46e-07/ 0. 4823	1.33e-06/ 0. 4280	1e-05/ 0.3815	
Sy- Sy-CelebA-0.5	1,100,682	3,005,533	6,330,995	200M
Sy-Sy-CelebA-0.7	96,836	339,832	894,335	200M
Sy- Sy-CelebA-1	5,521	22,921	73,003	200M
Sy-Sy-CelebA-HQ-0.5	1,813,871	4,299,750	8,119,721	200M
Sy-Sy-CelebA-HQ-0.7	209,072	641,038	1,499,136	200M
Sy- Sy-CelebA-HQ-1	13,004	50,335	146,326	200M
Sy- Sy-FFHQ-0.5	581,682	1,825,803	4,244,037	200M
Sy- Sy-FFHQ-0.7	32,633	124,973	365,651	200M
Sy- Sy- FFHQ-1	11,445	34,431	84,741	200M

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. Bazrafkan, H. Javidnia, 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 Understand.*, vol. 179, pp. 41–65, 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. 2818–2826.

[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] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," 2016, *arXiv:1612.02136*. [Online]. Available: http://arxiv.org/abs/1612.02136

[15] Z. Zhou, H. Cai, S. Rong, 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.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[19] Y. Ye, L. Wang, Y. Wu, Y. Chen, Y. Tian, Z. Liu, and Z. Zhang, "GAN quality index (GQI) by GAN-induced classifier," in *Proc. ICLR*, 2018, pp. 1–4.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[21] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," 2015, *arXiv:1511.06409*. [Online]. Available: http://arxiv.org/abs/1511.06409

[22] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3501–3510.

[23] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," 2015, *arXiv:1511.01844*. [Online]. Available: http://arxiv.org/abs/1511.01844

[24] A. Oliva, "Gist of the scene," in *Neurobiology of Attention*. Amsterdam, The Netherlands: Elsevier, 2005, pp. 251–256.

[25] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 15, pp. 6424–6429, Apr. 2007.

[26] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[27] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7064–7073.

[28] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley, "Semantically decomposing the latent spaces of generative adversarial networks," 2017, *arXiv:1705.07904*. [Online]. Available: http://arxiv.org/abs/1705.07904

[29] Y. Liu, Z. Qin, T. Wan, and Z. Luo, "Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks," *Neurocomputing*, vol. 311, pp. 78–87, Oct. 2018.

[30] Y. Lu, S. Wu, Y. W. Tai, C. K. Tang, and T. Youtu, "Sketch-to-image generation using deep contextual completion," 2017, *arXiv:1711.08972*. [Online]. Available: https://arxiv.org/abs/1711.08972

[31] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5077–5086.

[32] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[33] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "PacGAN: The power of two samples in generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1498–1507.

[34] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "VEEGAN: Reducing mode collapse in GANs using implicit variational learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3308–3318.

[35] S. Santurkar, L. Schmidt, and A. Madry, "A classification-based study of covariate shift in GAN distributions," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4480–4489.

[36] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[37] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2016, pp. 1–22.

[38] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 5040–5048.

[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[40] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6541–6549.

[41] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3387–3395.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*. [Online]. Available: http://arxiv.org/abs/1809.11096

[43] M. Marchesi, "Megapixel size image creation using generative adversarial networks," 2017, *arXiv:1706.00082*. [Online]. Available: http://arxiv.org/abs/1706.00082

[44] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10215–10224.

[45] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images, Detection, Alignment, Recognit.*, E. Learned-Miller, A. Ferencz, and F. Jurie, Eds., Marseille, France, Oct. 2008.

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[47] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," Tech. Rep., Aug. 2018, p. 2018, vol. 15.

[48] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: http://arxiv.org/abs/1710.10196

[49] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[50] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.

[51] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[52] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[53] A. Mansfield, *Information Technology–Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, document ISO/IEC 19795-1:2006, 2006, pp. 19791–19795.

[54] *ROC and AUC*, Google, Mountain View, CA, USA, 2020.
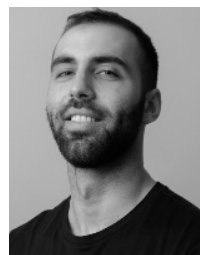
VIKTOR VARKARAKIS (Graduate Student Member, IEEE) received the B.Sc. degree in computer science and intelligent systems from the University of Piraeus, Greece, in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland Galway (NUIG). He is also with FotoNation/Xperi. His research interest includes machine learning using deep neural networks for tasks related to computer vision.

SHABAB BAZRAFKAN (Student Member, IEEE) received the B.Sc. degree in electrical engineering from Urmia University, Urmia, Iran, in 2011, the M.Sc. degree in telecommunication engineering, image processing branch from the Shiraz University of Technology (SuTECH), in 2013, and the Ph.D. degree in deep learning and neural network design from the National University of Ireland Galway (NUIG), in 2018. He is currently a Postdoctoral Researcher working on low dose CT image reconstruction using machine learning techniques with the Vision Laboratory, University of Antwerp.

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*.

● ● ●

Appendix G

Towards End-to-End Neural Face Authentication in the Wild – Quantifying and Compensating for Directional Lighting Effects.

Towards End-to-End Neural Face Authentication in the Wild – Quantifying and Compensating for Directional Lighting Effects.

Viktor Varkarakis[a], Wang Yao[a], Peter Corcoran[a]

[a]Department of Electronic Engineering, College of Engineering, National University of Ireland Galway, University Road, Galway, Ireland

Abstract

The recent availability of low-power neural accelerator hardware, combined with improvements in end-to-end neural facial recognition algorithms provides enabling technology for on-device facial authentication. The present research work examines the effects of directional lighting on a State-of-Art (SoA) neural face recognizer. A synthetic re-lighting technique is used to augment data samples due to the lack of public data-sets with sufficient directional lighting variations. Top lighting and its variants (top-left, top-right) are found to have minimal effect on accuracy, while bottom-left or bottom-right directional lighting have the most pronounced effects. Following the fine-tuning of network weights, the face recognition model is shown to achieve close to the original Receiver Operating Characteristic curve (ROC) performance across all lighting conditions, and demonstrates an ability to generalize beyond the lighting augmentations used in the fine-tuning data-set. This work shows that a SoA neural face recognition models can be tuned to compensate for directional lighting effects, removing the need for a pre-processing step prior to applying facial recognition.

Keywords: Directional Lighting, Face Illumination, Face Recognition, Face Re-Lighting Method

1. Introduction

Human Face Recognition (FR) has been an active research field in computer vision since the early 1990's [38] and early Convolutional Neural Network (CNN) based approaches were in evidence before the end of that decade

[19]. Over the last two decades FR has been well-studied in the literature with the most recent advances being driven by advances in CNN and deep learning [32, 36, 7, 41, 26]. In much of the literature the test samples for FR are assumed to be normalized in terms of pose, facial expression and illumination to simplify the challenge of accurately distinguishing an individual identity among a very large population. But, as it is not always feasible to capture optimal facial samples, some studies have explored the effects of different factors on the accuracy of State-of-Art (SoA) FR systems.

The main factors that affect FR include (i) pose [37, 47, 3, 18], (ii) illumination [2, 49, 42], (iii) facial expression [29, 30, 28], (iv) age [31, 6] and, (v) gender variations [44, 27].

In this work, the focus is on the latest end-to-end fully neural FR techniques [9] as these represent current SoA in terms of accuracy and have the potential for implementation in the latest neural accelerators [4, 12]. The initial focus for implementation of neural algorithms in embedded devices was on network optimizations such as parameter quantization and pruning, compressed convolutional filters and matrix factorizations [12]. However, the attention has recently shifted towards specialized neural topologies [22, 16] and ultra-low power realizations in hardware [10, 14]. Such optimizations enable a SoA neural FR architectures to be implemented in a low-power consumer appliance, enabling a new generation of devices capable of identifying their owners, provide access control and personalize the device's responses and behaviour. However, this introduces new challenges as such FR embodiment can no longer rely on pre-processing of input facial samples to optimize power consumption. Thus, all image processing must be achieved in a fully neural implementation, requiring a neural FR to be robust to factors such as pose and illumination. Here our goal is to determine the feasibility of modifying a high accuracy SoA neural FR architecture to demonstrate robustness to uncompensated input image samples.

This leads to the research questions posed in this work: (i) can we better quantify the effects of the external factors that affect fully neural FR and develop metrics to evaluate these; and (ii) can a fully neural FR architecture be modified through tuning and/or re-training to compensate internally for such external factors? As human interaction with a consumer device will typically lead the user to look directly at a screen, or similar user-interface, this work has a focus on the effects of lighting/illumination scenarios on FR. A consideration of the effects of facial pose and other external factors is left

2

to future studies. Specifically for the lighting variation, numerous image pre-processing methods exist to improve the performance of the FR model [35, 50, 34], but studies exploring the tuning or training of FR models to be robust to lighting variations are relatively rare [5, 21].

As a first step towards answering these research questions, this work employs a SoA re-lighting methodology to augment a set of high-quality facial images with directional lighting effects. The effect of these augmentations on the performance of a SoA FR method is quantified using Receiver Operating Characteristic curve (ROC) techniques. A similar approach was used recently to validate synthetic facial identities [40]. Note that a re-lighting augmentation approach was adopted as existing public datasets do not provide sufficient lighting variability. This is discussed in the sections 4 and 6. Finally, this works studies the feasibility of handling lighting variations by fine-tuning the neural FR network. The results for directional lighting are promising and indicate the potential for an end-to-end neural face authentication solution for in-the-wild faces.

2. Related Works

The advances in computational resources and with a surge in access to very large datasets, deep learning architectures have been developed and pushed the SoA in the FR task achieving exceptional accuracy results [26]. Famous deep neural approaches include DeepFace, FaceNet and ArcFace [32, 36, 7, 41]. Each work advanced the accuracy on the benchmarks of FR and new loss functions, pre-processing techniques and deep neural architectures were introduced. More information regarding the SoA of deep neural FR approaches as well as the entire pipeline of the FR and the methods used are given in [9, 1].

Despite the improvements, the FR task remained challenging in several cases. Studies revealed that many factors can have a negative impact on the FR performance, with the main factors being pose, illumination and others [2, 26, 18, 23]. Specifically, on face samples with lighting variation, techniques were proposed that compliment both the traditional and deep learning FR methods reporting improved performance [26, 15]. The approaches include pre-processing of the facial samples to normalise any variation, before feeding it to the face recognition algorithm [33, 17, 50, 34, 35].

Databases have been introduced to facilitate the development of FR models with light, pose, expression and other variations. The Yale b, CMU Multi-PIE, AR, Postech01 and UHDB31 [13, 25, 11, 8, 20] are a few of the datasets that have incorporate face variations and either focus on a single variation or a combination of different variations. Despite the developments of such datasets, the number of the variations and human subjects remain limited along with the fact that these datasets are usually acquired in controlled environments and therefore not being able to represent in-the-wild conditions.

Relighting techniques have been introduced in the literature with impressive results and able to introduce lighting into the face images without the degradation of the image by artifacts [48, 43] giving the ability to augment in-the-wild face datasets. Thus, providing a solution to the limited variation of lighting and human subjects in face datasets, which is discussed in section 4.

3. Methodology

In this section the various techniques and methodologies used in this work are detailed.

3.1. Face Re-Lighting Method

In this work, the lighting variation is applied to the CelebA-HQ dataset using the Deep Single Image Portrait Relighting (DPR) technique [48]. In this method a CNN is trained to generate a relighted image based on a Spherical Harmonics (SH) description of a lighting source. The method achieves SoA results, and in particular avoids introducing artifacts to the relighted samples - a drawback of other re-lighting methods that were considered for use in this study. The selected DPR method is trained on the well-known CelebA-HQ dataset which provides good variability in term of subject identity, combined with consistent face image quality. This makes the combination of the DPR re-lighting methodology and CelebA-HQ ideal for this work as side-effects are eliminated due to either variable facial sample quality or re-lighting artifacts, either of which could distort our experimental outcomes.

In this work we restricted our experiments to a select set of directional lighting components in order to gain a better understanding of the overall effect of directional lighting. The selected scenarios that are examined include lighting from 4 main directions: right, left, top and bottom of the face

image. This has the added benefit of keeping the computation requirements for experiments bounded to reasonable time-frame, with most individual experiments completing in less than a 48 hour period.

The representative Spherical Harmonic (SH) lighting sources used are shown in Fig.1. More SH lighting scenarios can be found in [1]. The illumination variations (right, left, top and bottom) are introduced to each sample of the CelebA-HQ dataset, resolving with 4 new sets of the CelebA-HQ, each containing of one illumination variation (CelebA-HQ-right, CelebA-HQ-left, CelebA-HQ-top and CelebA-HQ-bottom). Examples of the CelebA-HQ samples after introducing the illumination variations are illustrated in Fig.1. It can be seen from Fig.1 that the DPR method has high quality outputs incorporating the target SH lighting to the images realistically and without generating any artifacts to the face images. Instructions on how to generate the sets of CelebA-HQ with the different illumination scenarios are given in the Github repository of this work [2].

3.2. Face Recognition Model

A public reference implementation of the ArcFace [7] model is available, as the authors have released optimized, pre-trained, weights for the model.This reference ArcFace model has high performance on the dataset used in this work and provides a useful public baseline for future performance comparisons. This has motivated our use of ArcFace throughout this study. Other SoA FR models such as FaceNet [32] or CosFace [41] do not provide reference implementations and thus restrict direct experimental comparisons. An unofficial, but public, implementation of FaceNet [3] was also tested but could not provide a similar level of performance on the baseline or test datasets used in this work.

In this work, the recommended workflow, by the authors of ArcFace is followed before the face samples are fed to the network. The MTCNN [45] is used to detect and crop the face samples. The detected area is cropped and resized to 112×112, using bilinear interpolation, before passing to the ArcFace network which calculates the 512-embedding corresponding to a facial sample. Finally, the cosine similarity is used to compute the score

---

[1]https://zhhoper.github.io/dpr.html

[2]https://github.com/C3Imaging/Deep-Learning-Techniques/tree/Quantify-Retrain-FR-for-Light
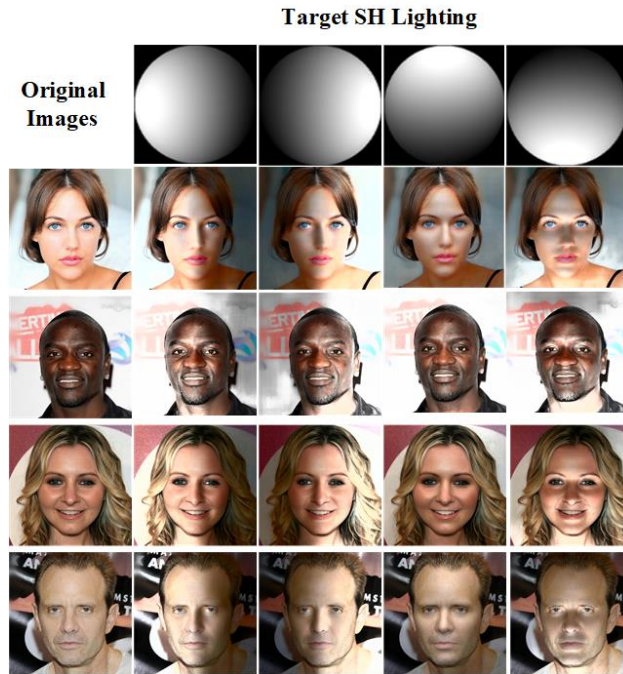
[3]https://github.com/davidsandberg/facenet

Figure 1: The target SH lighting that is examined in this work is presented on the top row. The original images of CelebA-HQ are on the left column. Examples of lighting injected in the original images using the DPR method [48] are shown for each examined illumination scenario (right, left, top & bottom).

representing the identity similarity when comparing two embeddings. The pretrained network used in this work is provided by the authors of ArcFace and can be found in [4].

Due to the introduction of the lighting variation and other factors, the face detection is not able to process all the face images from the CelebA-HQ sets. In the experiments only the images which the face detection network was able to process in all the illumination scenarios along with the original image, are used in order to keep the consistency in the experiments. Therefore, from the initial 30,000 images, in this work 28,222 are used from each set (CelebA-HQ-left, right, top, bottom and original). For the requirements of this study the dataset is divided into a train and test set with 19,570 images from 4k

identities and 8,654 images from 2k identities, respectively. This is applied to each CelebA-HQ-set. A list of the images used in the experiments can be found in [2].

3.3. Using ROC Curves as a Metric

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [24]. More specifically, to compute an ROC curve, an equal number of positive-identity and negative-identity pairs are created. Using the corresponding embeddings (extracted from the FR model for each image) of the pairs, similarity scores are calculated and used to plot the ROC curve. The closer an ROC curve is to unity, the better the performance of the FR model on the selected samples. More information regarding the ROC and its interpretation and use can be found in [40].

4. Quantifying the Effects of Illumination on the Face Recognition

In this section, the experiments conducted to quantify the effect of different lighting conditions on the FR's performance are presented with a discussion of the findings.

4.1. Experiments on Initial Lighting Variations

The effects of the 4 directional lighting scenarios shown in Fig.1 on the FR's performance are examined. Initially an ROC curve is calculated using only the samples from the test set of the original CelebA-HQ (ROC-Original) following the procedure described in section 3.3. All possible positive pairs from the test set of the original CelebA-HQ are used, in total 31k image pairs and an equal number of negative pairs are created randomly. Using the corresponding FR embeddings, the similarity scores are calculated and used to plot the ROC curve. The ROCs corresponding to re-lighting augmented scenarios are calculated following a similar procedure. The same positive and negative identity pairs as in ROC-Original are used but one of the samples from each pair has a re-lighting augmentation applied. Thus resulting in 4 main ROC curves (ROC-Left, ROC-Right, ROC-Top, ROC-Bottom) representing the FR's performance in each illumination scenario. The positive and negative pairs used to compute each ROC can be found in [2]. The resulting ROCs enable a direct comparison of the effects of different types of directional illumination with the original set of test image pairs and between

them. This is presented in Fig.2.

From Fig.2 the initial experimental results are largely self consistent and show well-defined performance degradation of the FR which is largely consistent with what might be expected. The ROC-Original curve illustrates that the FR model has a SoA performance on the non-augmented test dataset approaching close to unity, of 0.99 TPR on the corresponding to 10^{-4} FPR value. The re-lighting augmented ROC curves show significant deviations from this baseline performance and are largely consistent with what might be expected. Thus, the smallest deviation is for the ROC-Top, which starts at 0.925 TPR, followed by the ROC-Right and ROC-Left curves at 0.86 and 0.85 respectively. The worst performing ROC is that of the bottom light, starting a TPR of only 0.725. Looking at the examples shown in Fig.1 these results make sense - the top lighting augmentation causes the least distortion to the facial image from a human perspective, whereas the bottom-lighting creates more obvious distortions in the facial features. Finally, the left/right lighting augmentations would be expected to have similar effects due to the symmetry of a human face. Note that the slight variation between ROC-Left and ROC-Right is most likely due to slight left-right pose variations in some facial samples leading to eccentricities in the corresponding lighting augmentations.

4.2. Experiments on Additional Lighting Variations

The initial results shown in Fig.2 encouraged a more extensive set of experiments to include additional top-right, top-left and bottom-right, bottom-left lighting augmentations, to further improve our understanding of mixed directional lighting modes. The new SH lighting used and examples of the CelebA-HQ samples after introducing these illumination variations are illustrated in Fig.3. The goals of this additional set of experiments were to provide a second validation of our results, and in addition to explore the effects of more varied re-lighting augmentations.

Due to the introduction of the new lighting variations, the face detection is not able to process all the face images from the test set of the CelebA-HQ. Similarly as in section 3.2, only the images which the face detection network was able to process in all 8 illumination scenarios and the original images are used in order to keep the consistency in the experiments. Therefore, the initial test set of 8,654 images from 2k identities is reduced to 8,552 images from 1,979 identities. In order to calculate the ROCs, corresponding to the
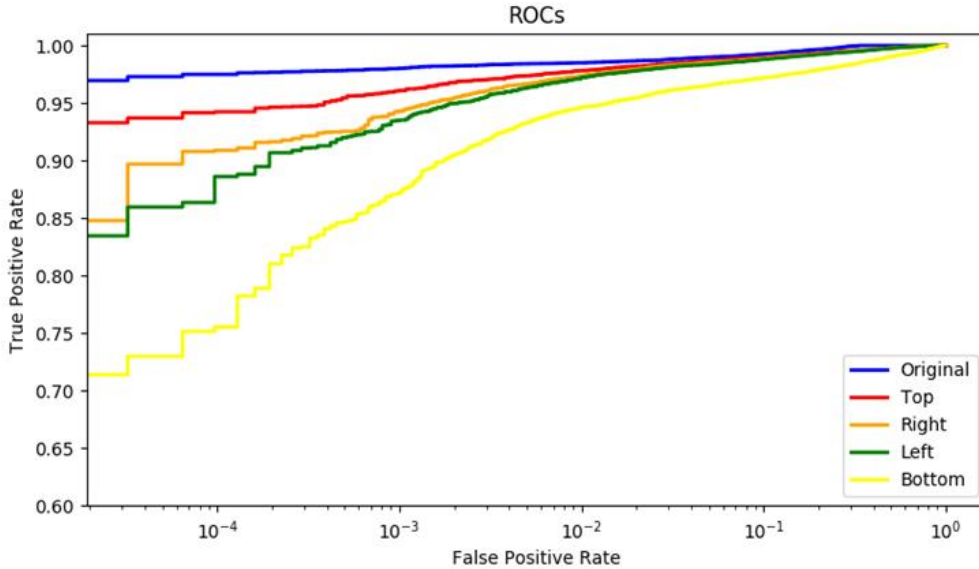
Figure 2: ROC curves, representing the performance of the FR model [4] on the original images and the 4 initial directional illumination scenarios (left, right, top, bottom) examined in this work.

original images and the 8 illuminations scenarios the procedure described in 3.3 and 4.1, is followed, using the new tests. As the size of the test set is reduced, so is the number of all possible positive image pairs used to compute the ROCs. For this set of experiments 30k positive pairs and an equal number of negative pairs are used. These pairs can be found in [2].

The primary directional ROCs curves (top, bottom, left & right) presented in Fig.4, differ slightly from those of Fig.2, as the image pairs used in these experiments are different. However their behaviour has a broadly similar characteristic. The bottom, bottom-left and bottom-right lighting augmentations are seen to be the most challenging for the FR task, while the top-left and top-right illuminations have the least effect on the FR's performance. There is a small drop and a small increase on the FR's performance from the top and right light respectively, compared to the Fig.2. These are attributed to the use of different pairs. These results are useful as a baseline for the section 5, as they help demonstrate that fine-tuning on the primary set of directional lighting augmentations can generalize across a broader range of directional lighting effects.
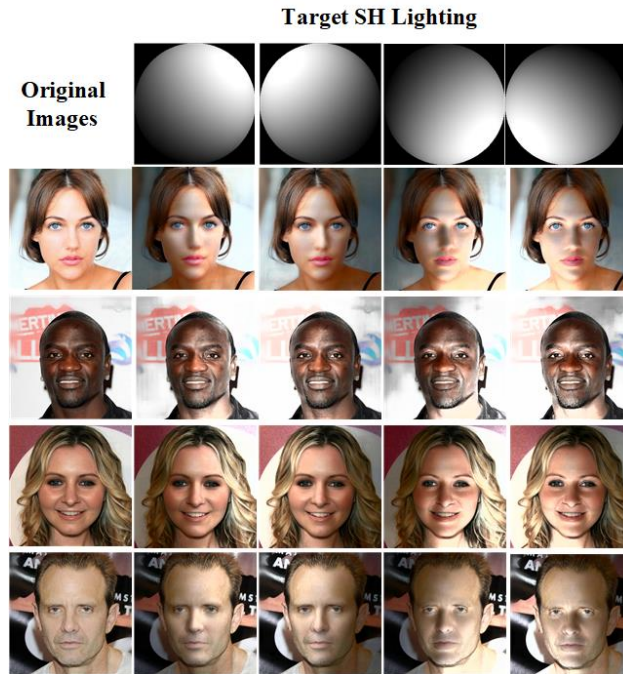
9

Figure 3: The additional SH lighting that is examined in section 4.2 is presented on the top row. The original images of CelebA-HQ are on the left column. Examples of lighting injected in the original images using the DPR method [48] are shown for each examined illumination scenario (top-left, top-right, bottom-left, bottom-right).

4.3. Experiments on Public Face Illumination Data-sets

Similar experiments as in 4.1 and 4.2, were conducted using face datasets which include illumination variation in their samples, to measure the effect of their lighting scenarios on the FR's performance. More specifically the AR and Postech01 [25, 8] face datasets were used. These experiments did not show any measurable degradation of the FR's performance across the illumination conditions provided in these data-sets. This is attributed to the relatively limited variation of the illumination scenarios, the numbers of human subjects, the total number of images and the controlled environments in which these face datasets were acquired. This is also indicated through the information regarding these datasets, provided in Table 1. Thus, revealing the need for the development of a face dataset that resembles in-the-wild conditions with illumination variation, which is further discussed in section 6.
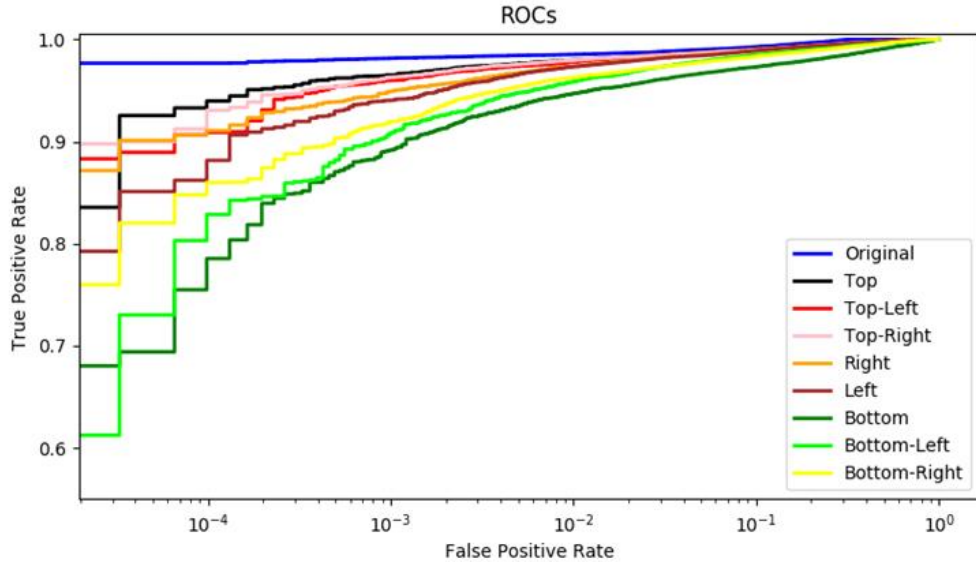
10

Figure 4: Additional ROC curves, representing the performance of the FR model [4] on the original images and the 8 directional illumination scenarios (left, top-left, bottom-left, right, top-right, bottom-right, top and bottom) examined in this work.

Table 1: Summary Information of Face Data-sets with Illumination Variations. (The number of Pose Variations in this table is referred with regards to the Illumination Variation)

Face Data-sets	#IDs	#Illumination Variations	#Pose Variations
AR [25]	126	3	1
Postech01 [8]	103	4	1

5. Fine-Tuning the Face Recognition model on Directional Illuminations

In the first part of this work the potential effects of in-the-wild lighting conditions, in particular directional lighting effects on a SoA neural face recognition method have been demonstrated and quantified. The next step is to determine whether the FR can be tuned to compensate for these effects. In this section the selected FR model is fine-tuned, using a similar approach to [39] with samples augmented with directional lighting.

11

5.1. Fine-tuning Process & ROCs Computation

The initial pretrained network provided by the authors of Arcface [4], is fine-tuned using a training set comprising samples from the original CelebA-HQ dataset and samples with all 4 primary directional lighting augmentations (CelebA-HQ-Left, Right, Top and Bottom). In total 97,850 high-quality facial samples were used for fine-tuning, or 19,750 from the original data and each of the four primary lighting sub-category.

For the re-training process the standard Arcface loss function is used, with the learning rate set to 0.005 and a batch size of 128, following the instructions from the authors of ArcFace. The network is fine-tuned for 40 epochs, as the number of images used is relatively large and all network layers are unfrozen for the fine-tuning process. After 40 epochs the network showed satisfactory results on the training data used and therefore stopped. Longer training could result in over-fitting to the training data and thus not being able to generalise. More details regarding the fine-tuning process and the corresponding training code can be found at [5]. The fine-tuned network resulting from this re-training process is released at [2].

The fine-tuned network, is used to calculate the embeddings of the test samples. The same procedure as described in 3.3 and 4.1 is followed to calculate the ROC-Original-finetuned (ROC-Original-FT) and the ROCs corresponding to the 8 different illumination scenarios (ROC-Left-FT, Right-FT, Top-FT, Bottom-FT, Top-Left-FT, Top-Right-FT, Bottom-Left-FT, Bottom-Right-FT,), using the positive and negative pairs from section 4.2. These ROCs are compared with the ROC-Original-FT and between them to explore whether the fine-tuned FR model is able to handle the variation in illumination as well as whether can generalise across the illuminations that were not used for the fine-tuning process.

5.2. ROCs Comparison

The ROCs representing the performance of the fine-tuned FR model on the original images and on the 8 directional lightings are presented in the Fig.5. From Fig.5 it is illustrated that the ROCs corresponding to the fine-tuned FR model on the 4 main illuminations (left, right, top, bottom) used in the fine-tuning process (Fig.5) are at higher levels compared to the ROCs
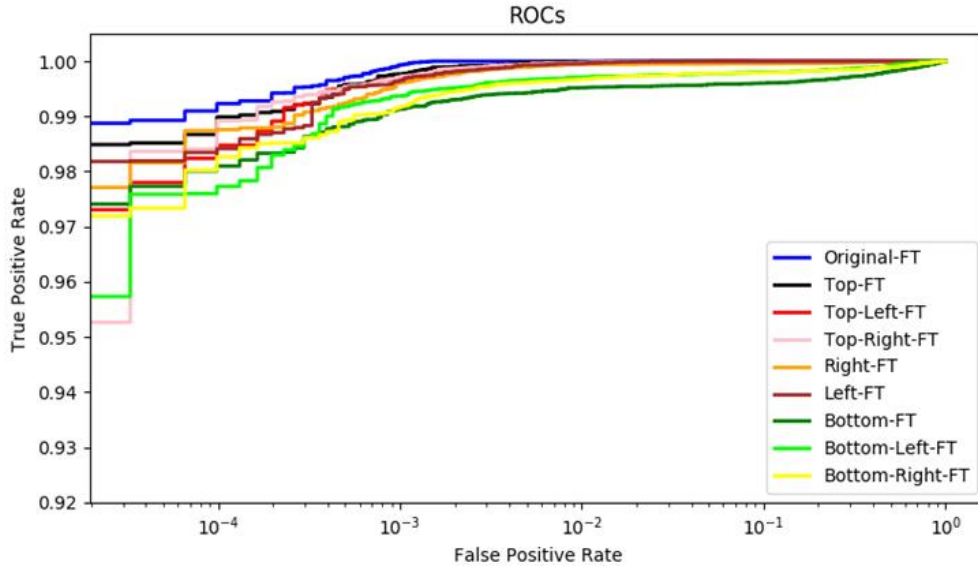
---

[5]https://github.com/deepinsight/insightface

Figure 5: ROC curves, representing the performance of the finetuned FR network [2] on the original images and the 8 directional illumination scenarios (left, top-left, bottom-left, right, top-right, bottom-right, top and bottom) examined in this work.

corresponding to the performance of the initial network (Fig.4) on these illumination scenarios. More importantly, the ROCs corresponding to the fine-tuned model on the 4 illuminations scenarios that are not used during the fine-tuning (top-left, top-right, bottom-left, bottom-right), are also at higher levels, thus showing that the network is able to generalise to other variations of illumination that it was not trained on. Overall, the performance of the fine-tuned FR model on any given illumination scenario has increased and its above 0.95 TPR, even on the lower FPR values. Notably, the ROCs are very close to the performance of the FR on the original images. Therefore, concluding that the FR model when trained with lighting variation is able to adopt and handle face samples that include illumination and achieve high accuracy results and also generalise across different illumination variations that are not used during fine-tuning. Thus, showing that the illumination can be compensated through training methods and augmentation techniques eliminating the need for pre-processing methods correcting the lighting, which are not optimal for use in neural accelerators.

6. Conclusion & Future Work

It is clear from the results of the experiments, illustrated in Fig. 2 & 4 that fully end-to-end neural FR solutions will be challenged by in-the-wild lighting conditions. As was indicated in [26] this problem is typically solved by additional pre-processing of image samples to correct for lighting conditions. In section 5.2 the practicality of fine-tuning a high-performing neural FR model has been demonstrated, recovering performance levels close to the original baseline for such lighting conditions. The fine-turning process also indicated that generalization from the primary directions to combinations of directional lighting is achieved - a promising result given the non-linear nature of lighting conditions. Note that providing a broader and more varied range of re-lighting samples and refining the training methodology to identify the more sensitive network layers in the ArcFace model should further improve these results, but even as they stand it is clear that a full end-to-end neural FR can be realized.

These initial results, especially the effectiveness of the fine-tuning process are very promising. They suggest that SoA neural FR algorithms can be fine-tuned to handle difficult in-the-wild acquisition conditions such as directional lighting. However there are other challenges for FR algorithms in-the-wild, including those listed in the introduction. A broader study on factors that can affect FR is indicated. In this regard the availability of several large 3D facial model datasets [46] could provide sufficient individual identities and support more complex data variations to support such a study.

References

[1] Adjabi, I., Ouahabi, A., Benzaoui, A., Taleb-Ahmed, A., 2020. Past, present, and future of face recognition: A review. Electronics 9, 1188.

[2] Beveridge, J.R., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Phillips, P.J., 2010. Quantifying how lighting and focus affect face recognition performance, IEEE. pp. 74–81.

[3] Choi, I.S., Choi, C.H., Kwak, N., 2011. Face recognition based on 2d images under illumination and pose variations. Pattern Recognition Letters 32, 561–571. doi:10.1016/j.patrec.2010.11.021.

[4] Corcoran, P., Lemley, J., Costache, C., Varkarakis, V., 2019. Deep learning for consumer devices and services 2—ai gets embedded at the edge. IEEE Consumer Electronics Magazine 8, 10–19. doi:10.1109/mce.2019.2923042.

[5] Crispell, D., Biris, O., Crosswhite, N., Byrne, J., Mundy, J.L., 2017. Dataset augmentation for pose and lighting invariant face recognition. arXiv preprint arXiv:1704.04326 .

[6] Deb, D., Best-Rowden, L., Jain, A.K., 2017. Face recognition performance under aging, pp. 548–556. doi:10.1109/CVPRW.2017.82.

[7] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition, pp. 4690–4699.

[8] Dong, H., Gu, N., 2001. Asian face image database pf01. 2006-01-05](2014-11). http://nava. postech. ac. kr/archives/irndb. html .

[9] Du, H., Shi, H., Zeng, D., Mei, T., 2020. The elements of end-to-end deep face recognition: A survey of recent advances. arXiv preprint arXiv:2009.13290 .

[10] Fleischer, B., Shukla, S., Ziegler, M., Silberman, J., Oh, J., Srinivasan, V., Choi, J., Mueller, S., Agrawal, A., Babinsky, T., Cao, N., Chen, C.Y., Chuang, P., Fox, T., Gristede, G., Guillorn, M., Haynie, H., Klaiber, M., Lee, D., Lo, S.H., Maier, G., Scheuermann, M., Venkataramani, S., Vezyrtzis, C., Wang, N., Yee, F., Zhou, C., Lu, P.F., Curran, B., Chang, L., Gopalakrishnan, K., 2018. A scalable multi-teraops deep learning processor core for ai trainina and inference, pp. 35–36. doi:10.1109/VLSIC.2018.8502276.

[11] Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J., 2001. From few to many: Illumination cone models for face recognition under variable

lighting and pose. IEEE transactions on pattern analysis and machine intelligence 23, 643–660.

[12] Goel, A., Tung, C., Lu, Y.H., Thiruvathukal, G.K., 2020. A survey of methods for low-power deep learning and computer vision, Institute of Electrical and Electronics Engineers Inc. doi:10.1109/WF-IoT48130.2020.9221198.

[13] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S., 2010. Multi-pie. Image and Vision Computing 28, 807–813.

[14] Guicquero, W., Verdant, A., 2020. Algorithmic enablers for compact neural network topology hardware design: Review and trends, pp. 1–5. doi:10.1109/iscas45731.2020.9181005.

[15] Han, H., Shan, S., Chen, X., Gao, W., 2013. A comparative study on illumination preprocessing in face recognition. Pattern Recognition 46, 1691–1699.

[16] He, Z., Gong, B., Fan, D., 2018. Optimize deep convolutional neural network with ternarized weights and high accuracy. arXiv .

[17] Hussain Shah, J., Sharif, M., Raza, M., Murtaza, M., Ur-Rehman, S., 2015. Robust face recognition technique under varying illumination. Journal of applied research and technology 13, 97–105.

[18] Lanitis, A., Taylor, C., Cootes, T., 1997. Automatic interpretation and coding of face images using flexible models. Pattern analysis and machine intelligence, IEEE Transactions on 19, 743–756. doi:10.1109/34.598231.

[19] Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D., 1997. Face recognition: A convolutional neural-network approach. IEEE Transactions on Neural Networks 8, 98–113. doi:10.1109/72.554195.

[20] Le, H.A., Kakadiaris, I.A., 2017. Uhdb31: A dataset for better understanding face recognition across pose and illumination variation, pp. 2555–2563.

[21] Le, H.A., Kakadiaris, I.A., 2019. Illumination-invariant face recognition with deep relit face images, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 2146–2155.

[22] Li, F., Zhang, B., Liu, B., 2016. Ternary weight networks .

[23] Mahmood, Z., Ali, T., Khan, S.U., 2016. Effects of pose and image resolution on automatic face recognition. IET biometrics 5, 111–119.

[24] Mansfield, A., 2006. Information technology–biometric performance testing and reporting–part 1: Principles and framework. ISO/IEC , 19791–19795.

[25] Martinez, A.M., 1998. The ar face database. CVC Technical Report24 .

[26] Mehdipour Ghazi, M., Kemal Ekenel, H., 2016. A comprehensive analysis of deep learning based representation for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 34–41.

[27] Narang, N., Bourlai, T., 2016. Gender and ethnicity classification using deep learning in heterogeneous face recognition. doi:`10.1109/ICB.2016.7550082`.

[28] Pavlović, M., Petrović, R., Stojanović, B., Stanković, S., 2018. Facial expression and lighting conditions influence on face recognition performance. Technical Report.

[29] Peña, A., Serna, I., Morales, A., Fierrez, J., Lapedriza, A., 2020. Facial expressions as a vulnerability in face recognition. arXiv preprint arXiv:2011.08809 .

[30] Peng, Y., Yin, H., 2018. Facial expression analysis and expression-invariant face recognition by manifold-based synthesis. Machine Vision and Applications 29, 263–284. doi:`10.1007/s00138-017-0895-6`.

[31] Riaz, S., Ali, Z., Park, U., Choi, J., Masi, I., Natarajan, P., 2019. Age-invariant face recognition using gender specific 3d aging modeling. Multimedia Tools and Applications doi:`10.1007/s11042-019-7694-1`.

[32] Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, pp. 815–823.

[33] Shan, S., Gao, W., Cao, B., Zhao, D., 2003. Illumination normalization for robust face recognition against varying lighting conditions, IEEE. pp. 157–164.

[34] Shinwari, A.R., Balooch, A.J., Alariki, A.A., Abdulhak, S.A., 2019. A comparative study of face recognition algorithms under facial expression and illumination, IEEE. pp. 390–394.

[35] Ruiz-del Solar, J., Quinteros, J., 2008. Illumination compensation and normalization in eigenspace-based face recognition: A comparative study of different pre-processing approaches. Pattern Recognition Letters 29, 1966–1979.

[36] Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: Closing the gap to human-level performance in face verification, pp. 1701–1708.

[37] Tran, L., Yin, X., Liu, X., 2017. Disentangled representation learning gan for pose-invariant face recognition, pp. 1283–1292. doi:`10.1109/CVPR.2017.141`.

[38] Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of cognitive neuroscience doi:`10.1162/jocn.1991.3.1.71`.

[39] Varkarakis, V., Bazrafkan, S., Corcoran, P., 2020a. Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets. Neural Networks 121, 101–121.

[40] Varkarakis, V., Bazrafkan, S., Costache, G., Corcoran, P., 2020b. Validating seed data samples for synthetic identities–methodology and uniqueness metrics. IEEE Access 8, 152532–152550.

[41] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018a. Cosface: Large margin cosine loss for deep face recognition, pp. 5265–5274.

[42] Wang, J.W., Le, N.T., Lee, J.S., Wang, C.C., 2018b. Illumination compensation for face recognition using adaptive singular value decomposition in the wavelet domain. Information Sciences 435, 69–93. doi:`10.1016/j.ins.2017.12.057`.

[43] Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F., 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Transactions on Graphics (TOG) 39, 1–13.

[44] Werther, C., Ferguson, M., Park, K., Kling, T., Chen, C., Wang, Y., 2018. Gender effect on face recognition for a large longitudinal database. arXiv .

[45] Xiang, J., Zhu, G., 2017. Joint face detection and facial expression recognition with mtcnn, IEEE. pp. 424–427.

[46] Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X., 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 601–610.

[47] Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., Yan, S., Feng, J., 2018. Towards pose invariant face recognition in the wild, pp. 2207–2216. doi:`10.1109/CVPR.2018.00235`.

[48] Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W., 2019. Deep single-image portrait relighting, pp. 7194–7202.

[49] Zhu, J.Y., Zheng, W.S., Lu, F., Lai, J.H., 2017. Illumination invariant single face image recognition under heterogeneous lighting condition. Pattern Recognition 66, 313–327. doi:`10.1016/j.patcog.2016.12.029`.

[50] Zou, X., Kittler, J., Messer, K., 2007. Illumination invariant face recognition: A survey, IEEE. pp. 1–8.

Appendix H

Versatile Auxiliary Classification and Regression With Generative Adversarial Networks

Versatile Auxiliary Classification and Regression With Generative Adversarial Networks

SHABAB BAZRAFKAN[1], (Student Member, IEEE),
VIKTOR VARKARAKIS[2], (Graduate Student Member, IEEE),
JOSEPH LEMLEY[2], (Member, IEEE), HOSSEIN JAVIDNIA[3],
AND PETER CORCORAN[2], (Fellow, IEEE)

[1]Department of Physics, University of Antwerp, 2000 Antwerp, Belgium
[2]College of Engineering and Informatics, National University of Ireland Galway, Galway, H91 TK33 Ireland
[3]School of Computer Science and Statistics ADAPT Centre, Trinity College Dublin, Dublin 2, D02 PN40 Ireland

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

ABSTRACT One of the most interesting challenges in Artificial Intelligence is to train conditional generators which are able to provide labeled adversarial samples drawn from a specific distribution. For a successful implementation of conditional generators, the created samples are constrained to a specific class. In this work, a new framework is presented to train a deep conditional generator by placing a classifier or regression model in parallel with the discriminator and back propagate the classification or regression error through the generator network. Special cases for binary classification, multi-class classification, and regression are studied. Experimental results on several data-sets are provided and the results are compared with similar state-of-the-art techniques. The main advantage of the method is that it is versatile and applicable to any variation of Generative Adversarial Network (GAN) implementation but also it is shown to obtain superior results compared to other methods. The mathematical proofs for the proposed scheme for both classification and regression are presented.

INDEX TERMS Conditional generators, deep neural networks, generative adversarial networks.

I. INTRODUCTION

Due in part to the affordability of modern parallel processing hardware, such as Graphical Processing Units (GPUs), the use of deep learning (DL) has become ubiquitous in solutions to a broad range of academic and industrial problems [1]. Deep Learning often provides superior outcomes on classification and regression problems compared to classical machine learning methods and other supervised learning techniques.

Deep learning has also proven successful in unsupervised learning such as with Generative Adversarial Networks (GAN) [2]. GANs, like other deep generative models, are able to learn an approximation of the distribution for a given task and generate an arbitrary number of samples from that distribution. These models comprise two networks, a generator, and a discriminator. The generator takes from a random latent vector, and the discriminator attempts to learn to distinguish

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

between images created by the generator (fake images) and those that are from the training set (real images). Since the original concept [2], many variations of GAN have been developed. Notable GANs include WGAN [3], EBGAN [4], BEGAN [5], DCGAN [6]. Other popular deep generative models include variational autoencoders, PixelRNN, and PixelCNN. Also, Conditional generative adversarial networks are models that can generate a class-specific data given the right latent input, such as CGAN [7]. Although these come with various disadvantages as some can be versatile to be used with any GAN structure but there is no mathematical proof showing that the trained generator is able to provide distinct samples for different classes [8]. Furthermore, other methods [8], [9], cannot be applied with any GAN structure. Additionally, the classification / regression term is restrained to the discriminator's structure methods [8], [9]. Thus, limiting the potential for improvements, as a more optimal classification/regression network cannot be select based on different tasks and relying on one structure for multiple problems.
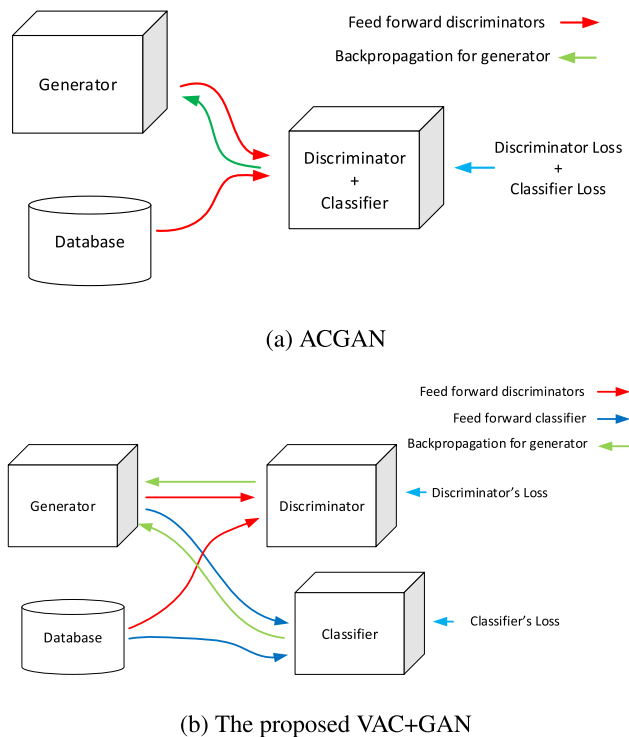
(a) ACGAN

(b) The proposed VAC+GAN

FIGURE 1. ACGAN [8] scheme vs VAC+GAN scheme.

The contribution of this work includes, two new training methods of Conditional GANs for classification and regression tasks, which improve the disadvantages discussed earlier. The first method is called Versatile Auxiliary Classifier with Generative Adversarial Network (VAC+GAN) applied for classification tasks. The initial idea of VAC+GAN is then extended to regression tasks. This approach is called Versatile Auxiliary Regression with Generative Adversarial Network (VAR+GAN). The main idea is to remove the classification/ regression term from the discriminator's loss function, by adding a classification /regression network that back-propagates through the generator. Also, in the VAR+GAN scheme a new loss function is also proposed. The main benefit of these methods is their versatility as they can be applied to any GAN structure with any loss function, as well as having the advantage of choosing any architecture for the classification/ regression network. Mathematical proofs are provided to show the applicability of the methods regardless of the GAN's and classifier's /regressor's structure or/and loss function. The proposed schemes are applied in the experiments section and compared with state-of-the-art methods, showing that using the proposed schemes resulted in improved performance.

The rest of the paper is arranged as follows: In section II related works are presented. In III, the mathematical proofs for both, VAC+GAN and VAR+GAN are given. In section IV, VAC+GAN is discussed and compared with state-of-the-art methods for binary and multi-class classification problems. VAR+GAN scheme is then discussed and compared with similar other state-of-the-art

techniques. Finally, the conclusions are presented in the last section (V).

II. RELATED WORKS

Conditional generative models are models that can generate a class-specific data given the right latent input. These have shown a significant improvement in generating good sample quality. Conditional GANs have shown a potential application for the image synthesis and image editing applications. Being able to augment a database for certain data classes/aspects and use them in training the final products is one of the most interesting applications for the conditional generators. In [7] the authors introduce a variation of GAN known as conditional GAN (CGAN), wherein the model is similar to the ordinary GAN [2], but the latent space is conditional with respect to the class label. This approach is versatile enough to be extended to other GAN variations, but there is no mathematical proof that the trained generator is able to provide distinct samples for different classes. A highly successful class-aware generative model is the Auxiliary Classifier GAN (ACGAN) [8], which is an extension of the CGAN [7] architecture. By adding a classification term to the generator and discriminator loss, the ACGAN's generator is forced to generate a specific class of data for a given input (See Fig. 1a). One disadvantage of ACGAN [8] is that it cannot be used with arbitrary GANs. Mixing the loss of discriminator and the classifier will alter the training convergence if the output of the discriminator is from a different type compare to the classifier's output. Furthermore, Information maximizing Generative Adversarial Network (Info-GAN) [10] splits an input latent space into the standard noise vector z and additional latent vector c. The latent vector c is then made meaningful disentangled representation by maximizing the mutual information between latent vector c and generated images $G(z, c)$ using additional Q network. Also, the Similarity constraint Generative Adversarial Network (SC-GAN) [11] attempts to learn disentangled latent representation by adding the similarity constraint between latent vector c and generated images $G(z, c)$. Info-GAN [10] uses an extra network to learn disentangle representation, while SCGAN only adds an additional constraint to a standard GAN. Therefore, SCGAN [11] simplifies the architecture of Info-GAN. Finally, related work includes [9] wherein a Hierarchical Generative Model (HGM) is utilized for eye image synthesis and eye gaze estimation. This work introduces a variation of GAN known as conditional Bidirectional GAN (cBiGAN) [9], shown in Fig. 2a, which is a mixture of CGAN and Bidirectional GAN (BiGAN). The primary disadvantage with this method is that it can only be used with BiGAN like networks.

III. MATHEMATICAL PROOFS

In this section, the mathematical proofs of the proposed techniques (VAC+GAN and VAR+GAN) are given, showing their applicability and effectiveness to any GAN implementation using any classification/regression network.
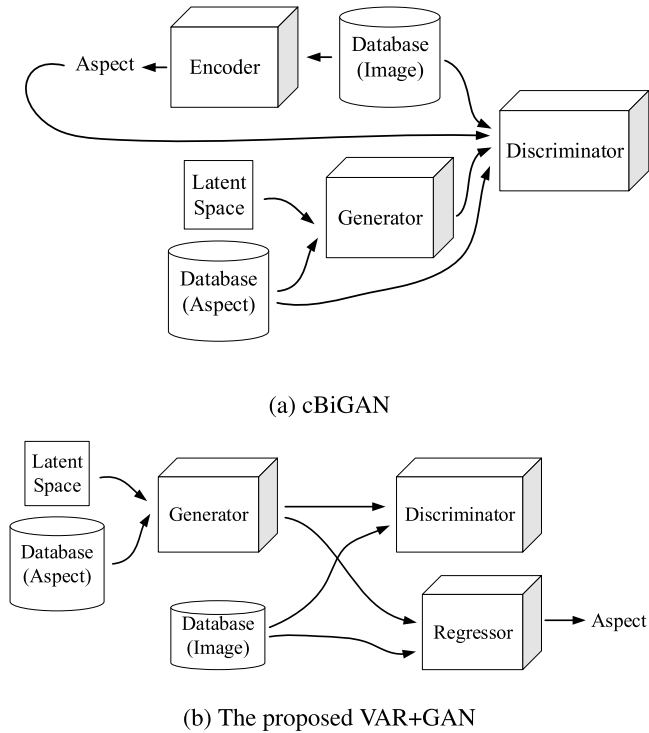
(a) cBiGAN

(b) The proposed VAR+GAN

FIGURE 2. cBiGAN [9] scheme vs VAR+GAN scheme.

A. VERSATILE AUXILIARY CLASSIFIER WITH GENERATIVE ADVERSARIAL NETWORK (VAC+GAN)

The proposed method of VAC+GAN, combines a GAN and a classifier in such a way that the classifier accepts samples from the generator, and the classification error is back-propagated through the classifier and the generator. The flowchart of the proposed model's structure is presented in Fig. 1b.

In this section it is shown that by placing a classifier at the output of the generator and minimizing the categorical cross-entropy as the classifiers loss, the Jensen-Shannon Divergence between all the classes is increased. The terms used in the mathematical proofs are the following:

1) N is the number of the classes.
2) The latent space Z is partitioned in to $\{Z_1, Z_2, \ldots, Z_N\}$ subsets. This means that $\{Z_1, Z_2, \ldots, Z_N\}$ are disjoint and their union is equal to the Z-space.
3) C is the classifier function.
4) $\mathcal{L}_{ce}$ is the binary cross-entropy loss function.
5) $\mathcal{L}_{cce}$ is the categorical cross-entropy loss function.

Proposition 1: In the multiple classes case, the classifier C has N outputs, where N is the number of the classes. In this approach, each output of the classifier corresponds to one class. For a fixed Generator and Discriminator (meaning the weights of the discriminator and generator do not change but the classifier's weights do change), the optimal output for class c (c'th output) is:

$$C^*_{G,D}(c) = \frac{p_{X_c}(x)}{\sum_{i=1}^{N} p_{X_i}(x)} \quad (1)$$

Proof: Considering just one of the outputs of the classifier, the categorical cross-entropy can be reduced to binary cross-entropy given by

$$\mathcal{L}_{ce}(C(c)) = -\mathbb{E}_{z \sim p_{Z_c}(z)}\big[\log\big(C(G(z))\big)\big]$$
$$-\mathbb{E}_{z \sim \sum_{i \neq c} p_{Z_i}(z)}\big[1 - \log\big(C(G(z))\big)\big] \quad (2)$$

which is equal to:

$$\mathcal{L}_{ce}(C(c)) = \int \Big(p_{Z_c}(z) \log\big(C(G(z))\big)$$
$$+ \big(\sum_{i \neq c} p_{Z_i}(z)\big) \log\big(1 - C(G(z))\big) dz \Big) \quad (3)$$

By considering $G(z_i) = x_i$:

$$\mathcal{L}_{ce}(C(c)) = \int \Big(p_{X_c}(x) \log(C(x))$$
$$+ \big(\sum_{i \neq c} p_{X_i}(x)\big) \log(1 - C(x)) dx \Big) \quad (4)$$

The function $f \to m \log(f) + n \log(1 - f)$ gets its maximum at $\frac{m}{m+n}$ for any $(m, n) \in \mathbb{R}^2 \setminus \{0, 0\}$, concluding the proof. □

Theorem 1: The maximum value for $\mathcal{L}_{cce}(C)$ is $N \log(N)$ and is achieved if and only if $p_{X_1} = p_{X_2} = \ldots = p_{X_N}$.

Proof: The categorical cross-entropy is given by:

$$\mathcal{L}_{cce} = -\sum_{i=1}^{N} \mathbb{E}_{z \sim p_{Z_i}(z)}\big[\log\big(C(G(z))\big)\big]$$
$$= -\sum_{i=1}^{N} \int p_{X_i}(x) \log(C(x)) dx \quad (5)$$

From equation 1:

$$\mathcal{L}_{cce} = -\sum_{i=1}^{N} \left(\int p_{X_i}(x) \log\left(\frac{p_{X_i}(x)}{\sum_{j=1}^{N} p_j(x)}\right) dx \right)$$
$$= N \log(N)$$
$$-\sum_{i=1}^{N} \left(\int p_{X_i}(x) \log\left(\frac{p_{X_i}(x)}{\sum_{j=1}^{N} \frac{p_j(x)}{N}}\right) dx \right)$$
$$= N\log(N) - \sum_{i=1}^{N} KL\left(p_{X_i}(x) \bigg\| \sum_{j=1}^{N} \frac{p_{X_j}(x)}{N}\right) \quad (6)$$

where KL is the Kullback-Leibler divergence, which is always positive or equal to zero.

Now consider $p_{X_1} = p_{X_2} = \cdots = p_{X_N}$. From 6:

$$\mathcal{L}_{cce} = N \log(N) - \sum_{i=1}^{N} KL\left(p_{X_i}(x) \bigg\| p_{X_i}(x)\right)$$
$$= N \log(N) \quad (7)$$

concluding the proof. □

Theorem 2: Minimizing $\mathcal{L}_{cce}$ increases the Jensen-Shannon Divergence between $p_{X_1}, p_{X_2}, \ldots, p_{X_N}$

Proof: From equation 6:

$$\mathcal{L}_{cce} = N \log(N) - \int \sum_{i=1}^{N} \Big(p_{X_i}(x)\Big[\log(p_{X_i}(x))$$
$$- \log\Big(\sum_{j=1}^{N} \frac{p_{X_j}(x)}{N}\Big)\Big]\Big) dx \quad (8)$$

Which can be rewritten as:

$$\mathcal{L}_{cce} = N\log(N) - \sum_{i=1}^{N}\left(\int p_{X_i}(x)\log(p_{X_i}(x))dx\right)$$
$$+ \underbrace{\int \sum_{i=1}^{N}\left(p_{X_i}(x)\log\left(\sum_{j=1}^{N}\frac{p_{X_j}(x)}{N}\right)\right)dx}_{\left(\sum_{i=1}^{N}p_{X_i}(x)\right)\left(\log\left(\sum_{j=1}^{N}\frac{p_{X_j}(x)}{N}\right)\right)} \quad (9)$$

Which is equal to:

$$\mathcal{L}_{cce} = N\log(N)$$
$$- N\sum_{i=1}^{N}\left(\frac{1}{N}\int p_{X_i}(x)\log(p_{X_i}(x))dx\right)$$
$$+ N\int\left(\sum_{i=1}^{N}\frac{p_{X_i}(x)}{N}\right)\left(\log\left(\sum_{j=1}^{N}\frac{p_{X_j}(x)}{N}\right)\right)dx \quad (10)$$

This equation can be rewritten as:

$$\mathcal{L}_{cce} = N\log(N)$$
$$- \left[H\left(\sum_{i=1}^{N}\frac{1}{N}p_{X_i}(x)\right) - \sum_{i=1}^{N}\frac{1}{N}H\left(p_{X_i}(x)\right)\right] \quad (11)$$

wherein the $H(p)$ is the Shannon entropy of the distribution p.

The Jensen Shannon divergence between N distributions $p_1, p_2, \ldots, p_N$, is defined as:

$$JSD_{\pi_1,\pi_2,\ldots,\pi_N}\left(p_1, p_2, \ldots, p_N\right) = H\left(\sum_{i=1}^{N}\pi_i p_i\right)$$
$$- \sum_{i=1}^{N}\pi_i H(p_i) \quad (12)$$

From equations 11 and 12:

$$\mathcal{L}_{cce} = N\log(N) - N\,JSD_{\frac{1}{N},\frac{1}{N},\ldots,\frac{1}{N}}$$
$$\left(p_{X_1}(x), p_{X_2}(x), \ldots, p_{X_N}(x)\right) \quad (13)$$

Minimizing $\mathcal{L}_{cce}$ increases the JSD term, concluding the proof. □

In this section it has been shown that by placing a classifier at the output of the generator and back-propagate the classification error throughout the generator, one can increase the dis-similarity between the classes for the generator. Therefore deep generator can be trained which can produce class specified samples. The mathematical proofs are presented for a multi-class scenario but consequently the technique works for binary classification also by setting N (number of classes) equal to 2.

B. VERSATILE AUXILIARY REGRESSOR WITH GENERATIVE ADVERSARIAL NETWORK (VAR+GAN)

In the previous subsection, mathematical proof for the applicability and effectiveness of VAC+GAN, which deals with classification tasks, was presented. In this subsection a similar proof for the case of regression tasks, called VAR+GAN is given. VAR+GAN, uses a regression network in addition to the usual discriminator and generator network and the error from the regressor network back propagates to the generator. The flowchart of the proposed VAR+GAN is presented in Fig. 2b)

In this method, the generator is constrained to generate samples with specific continuous aspects. Aspects refers to the input of the generator and the output of the regressor. For example, in the face generation task, given the right latent sequence, the generator is able to create faces with particular landmarks. The following loss function is introduced for the regression network:

$$L_R = \int\int dp_z(z)\left(-\log\left(1 - (y - R(G(z)))\right)\right)dz \quad (14)$$

wherein z is the latent space variable, $dp_z(z)$ is the distribution of an infinitesimal partition of latent space, y is the target variable (ground truth), R is the regression function and G is the generator function.

Proposition 2: For the loss function in equation 14 the optimal regressor is:

$$R^* = \frac{p(x)}{c} + y - 1 \quad (15)$$

wherein $p(x)$ is the distribution of the generator's output, c is post-integration constant, and y is the target function.

Proof: Considering the inner integration of equation 14 and by replacing $G(z) = x$, the extremum of the loss function with respect to R is:

$$\frac{d}{dR}\int dp_x(x)\left(-\log\left(1 - (y - R(x))\right)\right)dx = 0 \quad (16)$$

which can be written as:

$$\int\frac{-dp_x}{R - y + 1} = 0 \quad \Rightarrow \quad \frac{p_x}{R - y + 1} = c \quad (17)$$

this results in:

$$R = \frac{p(x)}{c} + y - 1 \quad (18)$$

concluding the proof. □

Theorem 3: Minimizing the loss function in equation 14 decreases the entropy of the generator's output.

Proof: by replacing the equation 15 in 14, gives :

$$L_R = \int\int -\log\left(\frac{p_x(x)}{c}\right)dp_x dx \quad (19)$$

which can be rewritten as:

$$L_R = -\int p_x(x)\log(p_x(x))dx + \log(c)$$
$$= H(p_x(x)) + \log(c) \quad (20)$$

wherein H is the shannon entropy. Minimizing L_R results in decreasing $H(p_x(x))$ concluding the proof. $\qquad\square$

Adding the regressor to the model decreases the entropy of the generated samples. This is expected since the idea is to constrain the output of the generator to obey some particular criteria. This is shown in observations in section IV-C4.

Theorem 4: For any two sets of samples and their corresponding targets (y_1 and y_2), the loss function in equation 14 increases the Jensen Shannon Divergence (JSD) between generated samples for these two sets.

Proof: Consider z_1 and z_2 are two partitions of the latent space correspond to two sets of samples with targets y_1 and y_2. In this case, the loss function in equation 14 is given by:

$$L_R = -\int p_{z_1}(z) \log(1 - (y_1 - R(G(z_1)))) dz$$
$$- \int p_{z_2}(z) \log(1 - (y_2 - R(G(z_2)))) dz \quad (21)$$

Considering $G(z_1) = x_1$, $G(z_2) = x_2$, $c_1 = 1 - y_1$, and $c_2 = 1 - y_2$ equation 21 simplifies to:

$$L_R = -\int p_{x_1}(x) \log(c_1 + R(x)) dx$$
$$- \int p_{x_2}(x) \log(c_2 + R(x)) dx \quad (22)$$

To find the optimum $R(x)$ the derivative of the integrand is set to zero given by:

$$\frac{p_{x_1}}{c_1 + R} + \frac{p_{x_2}}{c_2 + R} = 0 \quad (23)$$

which results in

$$R = -\frac{p_{x_1} c_2 + p_{x_2} c_1}{p_{x_1} + p_{x_2}} \quad (24)$$

By replacing equation 24 in equation 22 it simplifies to:

$$L_R = -\int p_{x_1} \log\left(\frac{(c_1 - c_2) p_{x_1}}{p_{x_1} + p_{x_2}}\right)$$
$$+ \int p_{x_2} \log\left(\frac{(c_2 - c_1) p_{x_2}}{p_{x_1} + p_{x_2}}\right) dx \quad (25)$$

which can be rewritten as:

$$R_L = -\int \log\left(\frac{p_{x_1}}{\frac{p_{x_1} + p_{x_2}}{2}}\right) - \int \log\left(\frac{p_{x_2}}{\frac{p_{x_1} + p_{x_2}}{2}}\right)$$
$$- \log(c_1 - c_2) - \log(c_2 - c_1) - \log(4) \quad (26)$$

which equals to:

$$R_L = -\log(c_1 - c_2) - \log(c_2 - c_1) - \log(4)$$
$$- 2JSD(p_{x_1} || p_{x_2}) \quad (27)$$

minimizing R_L increases $JSD(p_{x_1} || p_{x_2})$ term, concluding the proof. $\qquad\square$

In this section, it has been shown, that the presented loss function increases the distance between generated samples for any two set of aspects. This is a desirable feature because even small changes in the generator's input space causes changes in the generated images. This protects against common problems such as mode collapse or near mode collapse and increases the diversity of the generated samples.

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

In this section the implementation of VAC+GAN and VAR+GAN is described. In continuance, experimental details are provided and the proposed methods are compared with similar state-of-the-art techniques. Specifically, experiments are conducted for three main scenarios: VAC+GAN for binary and multi-class classification and the VAR+GAN for regression. The code for both methods can be in [1] and.[2]

A. BINARY VAC+GAN

The VAC+GAN method is implemented using the BEGAN [5] structure for training a gender specified generator on the CelebA [12] database. The results of the proposed method are compared against the results of the conditional GAN (CGAN) [7] method applied also to the BEGAN [5] structured, for a fair comparison on the task of generating face samples constrained to either male or female class.

Different metrics are used to show the diversity of the generated samples including Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Universal Quality Index (UQI), and Structural Similarity Index (SSIM). These metrics are explained in Appendix. All the networks mentioned are trained with Lasagne [13] on top of the Theano [14] library in Python.

A comparison with the ACGAN method is not available since applying this method to the BEGAN framework resulted in mode collapse on the first epoch.

1) DATABASE

The CelebA dataset [12] consists of 202,599 original samples. The OpenCV frontal face cascade classifier [15] is utilised to detect facial regions which are cropped and resized to 48×48 pixels and used for training the proposed GAN framework.
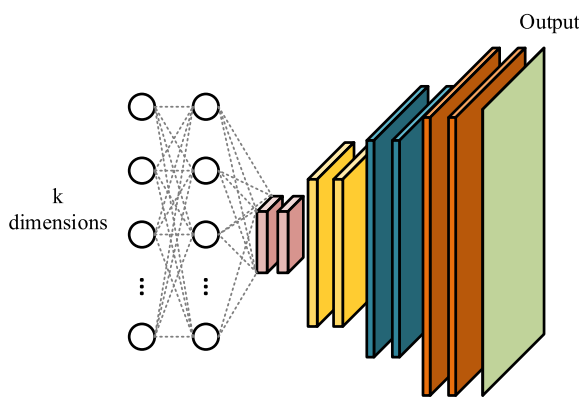
2) NETWORK ARCHITECTURES

In the BEGAN [5] structure, the generator network it has the same architecture as the decoder part of an auto-encoder. The network used in the experiment contains one fully connected layer which reshapes the input to match next layer. The next layers are all convolutional layers followed by $(2, 2)$ un-pooling layers for every second convolution. The exponential linear unit (ELU) [16] is used as activation function except in the last layer wherein no non-linearity has been applied. The discriminator network used in this experiment is an auto-encoder. The input of the auto-encoder is the image (48×48). The encoder part of the network is made of convolutional layers with ELU activation function. The down-scaling in these layers is obtained by using $(2, 2)$ stride in every second convolutional layer. The architecture of decoder is the same as the generator network. The bottleneck of the auto-encoder is a fully connected layer with no activation function. The encoder and decoder networks

---

[1]https://github.com/shababqcd/VAC-GAN
[2]https://github.com/shababqcd/VAR-GAN
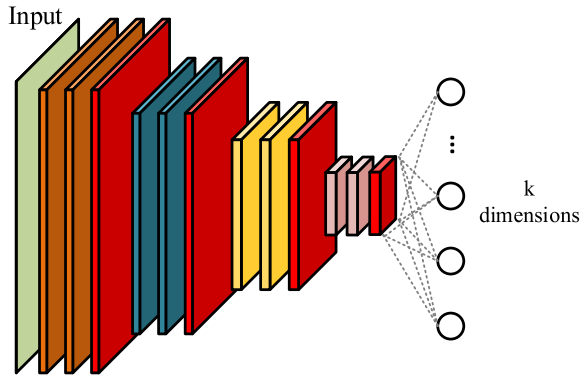
(a) Encoder network.

(b) Decoder network.

FIGURE 3. Network architectures used for implementation purposes in the experiments of IV-A and IV-C.

TABLE 1. The classifier structure for the experiment in IV-A.

Layer	Type	kernel	Activation
Input	Input(48×48)	–	–
Hidden 1	Conv	3×3(16 ch)	ReLU
Pool 1	Max pooling	2×2	–
Hidden 2	Conv	3×3(8 ch)	ReLU
Pool 2	Max pooling	2×2	–
Hidden 3	Dense	1024	ReLU
Output	Dense	1	Sigmoid

used for training the BEGAN [5] are shown in Fig. 3a and 3b respectively. The layers shown in red apply no non-linearity to the data. The classifier used in this experiment to implement the VAC+GAN proposed scheme is given in Table 1.

3) IMPLEMENTATION
a: VAC+GAN
The loss functions used to train VAC+GAN applied to BEGAN [5] framework are given by

$$L_d = L(x) - k_t \cdot L(G(z|c))$$
$$L_g = \vartheta \cdot L(G(z|c)) + \zeta \cdot BCE$$
$$k_{t+1} = k_t + \lambda_k \big(\gamma L(x) - L(G(z|c))\big) \qquad (28)$$

where L_g and L_d are the generator and discriminator losses respectively. G is the generator function, z is a sample from the latent space, c is the class label, x is the sample drawn from the database, λ_k is the learning rate for k, γ is the equilibrium hyper parameter set to 0.5 in this work, and L is the auto-encoders loss defined by

$$L(v) = |v - D(v)|^2 \qquad (29)$$

BCE is the binary cross-entropy loss of the classifier, and ϑ and ζ are set to 0.997 and 0.003 respectively. The optimizer used for training the generator and discriminator is ADAM with learning rate, β_1 and β_2 equal to 0.0001, 0.5 and 0.999 respectively. The classifier is optimized using nestrov momentum gradient descent with learning rate and momentum equal to 0.01 and 0.9 respectively.

The latent space has 64 dimensions and the first dimension is used to partition the latent space in two subspaces corresponding to the two classes.

b: CBEGAN
The loss functions for training the conditional BEGAN (CBEGAN) [5] are given by:

$$L_d = L(x) - k_t \cdot L(G(z|c))$$
$$L_g = L(G(z|c))$$
$$k_{t+1} = k_t + \lambda_k \big(\gamma L(x) - L(G(z|c))\big) \qquad (30)$$

where L_g and L_d are the generator and discriminator losses respectively. G is the generator function, z is a sample from the latent space, c is the class label, x is the sample drawn from the database, λ_k is the learning rate for k, γ is the equilibrium hyper parameter set to 0.5 in this work, and L is the auto-encoders loss defined in equation 29.

4) RESULTS
The results from experiments involving CBEGAN [5] and the proposed VAC+GAN method are shown in Fig. 4 and 5, respectively.

Fig. 4 shows that the gender-specific generator fails to correctly generate samples for a specific class when the conditional GAN is applied (CBEGAN) [5]. The Fig. 5 indicates that the proposed VAC+GAN method constrains the generator to create samples drawn from a specific class.

In order to compare the models (CBEGAN [5], VAC+GAN), 80 random male and 80 random female samples have been generated using the trained generators and three observations have been conducted on these samples:

1) Each male sample has been compared to all the other male samples, and all the metrics have been calculated for these comparisons, and the average of these numbers has been obtained (blue bars).
2) Each female sample has been compared to all the other female samples, and all the metrics have been calculated for these comparisons, and the average of these numbers has been obtained (purple bars).
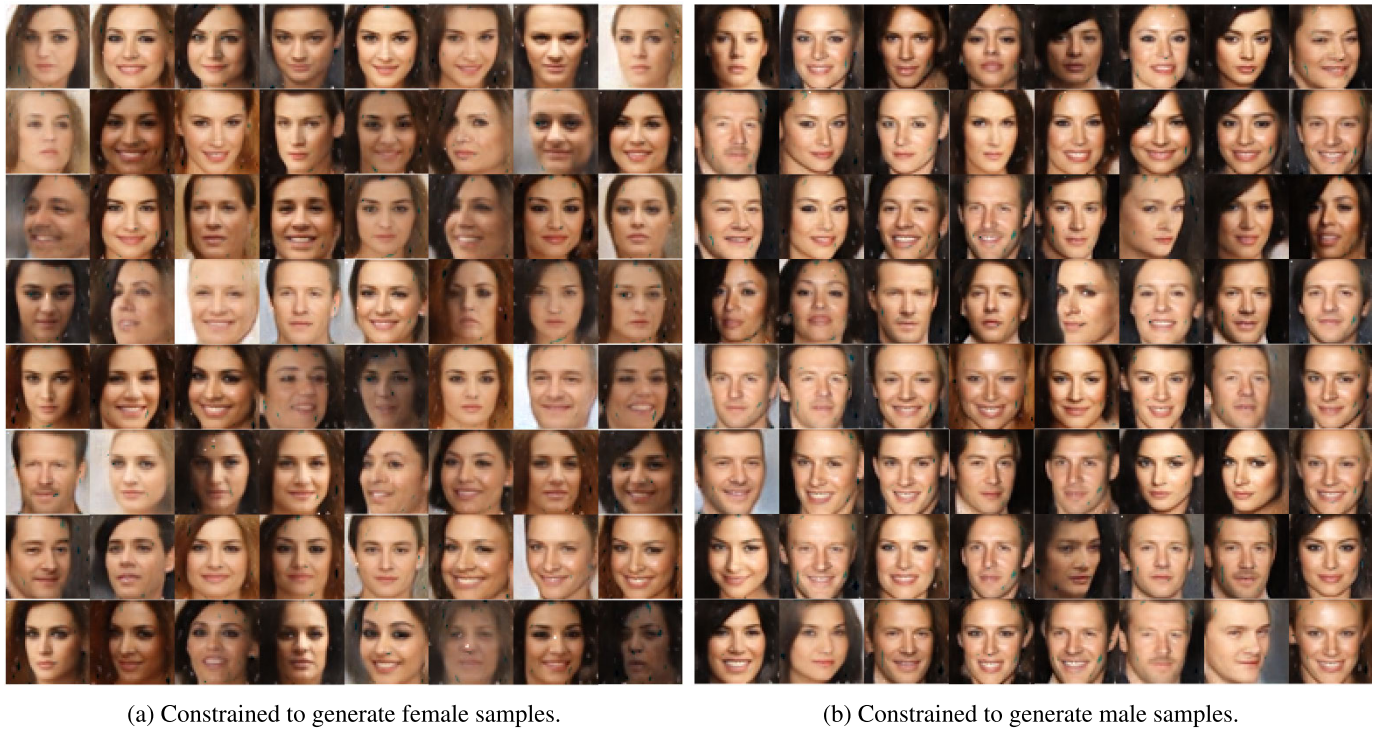
(a) Constrained to generate female samples.

(b) Constrained to generate male samples.

FIGURE 4. Generator trained using CBEGAN [5] method.

(a) Constrained to generate female samples.
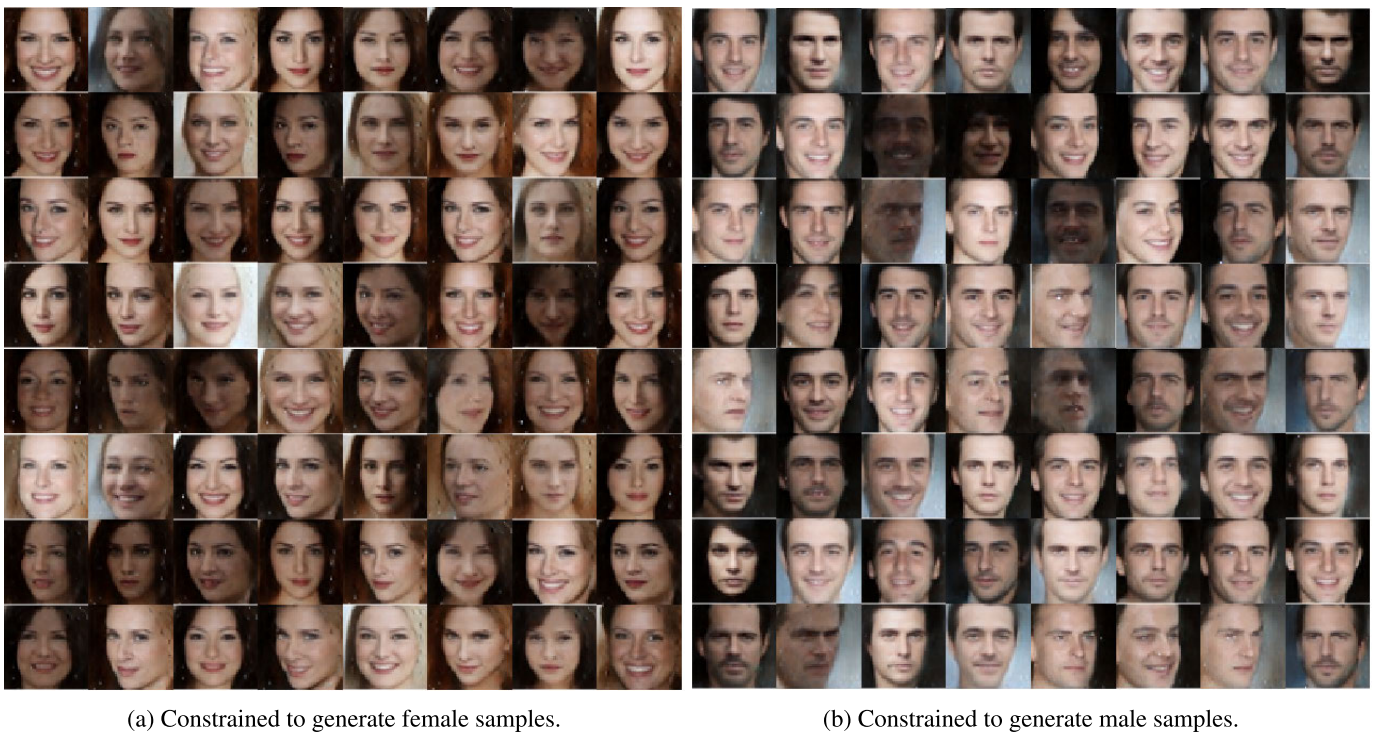
(b) Constrained to generate male samples.

FIGURE 5. Generator trained using the proposed VAC+GAN method.

3) Each male sample has been compared to all female samples, and all the metrics have been calculated for these comparisons, and the average of these numbers has been obtained (yellow bars).

The aforementioned measurements are illustrated in Fig. 6 and 7 for the CBEGAN [5] and the proposed method. The lower value of UQI and SSIM show low similarity between samples. In Fig. 6, from the first two observations (blue
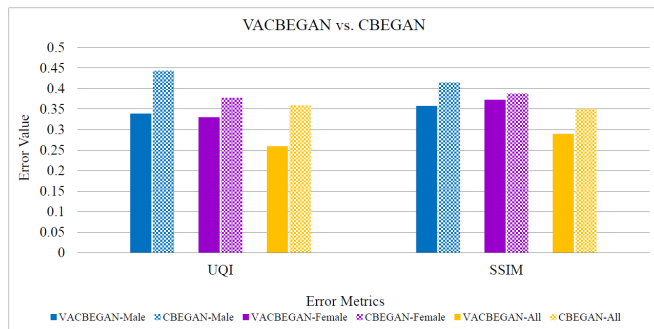
FIGURE 6. UQI and SSIM metrics for VAC+GAN vs. CBEGAN [5]. Lower values show higher performance.
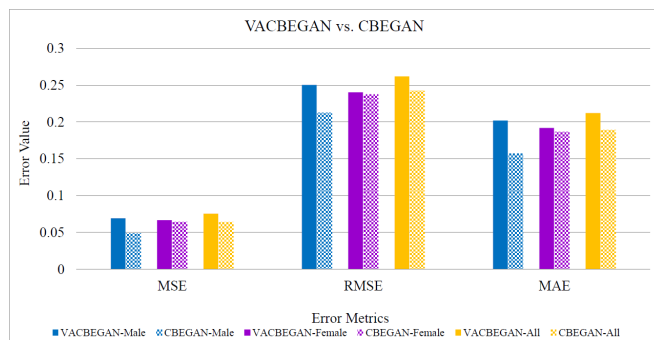

FIGURE 7. MSE, RMSE and MAE metrics for VAC+GAN vs. CBEGAN [5]. Higher values show better performance.

and purple bars) it is shown that the proposed method is able to draw samples from each class that are not similar. From the third observation (yellow bars), it is shown that the inter-class similarity in the proposed method is less than the one from CBEGAN [5]. This demonstrates that the generated samples of different classes from the proposed VAC+GAN method differ more compared to the same measure applied to CBEGAN [5] approach. Finally, the higher value of MSE, RMSE, and MAE show high variation of the generated images for the proposed method. As shown in Fig. 7, the proposed method is able to generate samples with high variation for each class and also between classes.

B. MULTI-CLASS VAC+GAN

In this section, two experiments are conducted to show the effectiveness of VAC+GAN in multi-class classification scenarios. In the first one, the VAC+GAN scheme is applied to the DCGAN [6] structure and trained on the MNIST [17] dataset. Visual comparisons are conducted with CGAN [7],CDCGAN [6] and ACGAN [8] on the task of generating MNIST samples constrained to a class.

The second experiment is implemented to the ACGAN [8] structure and compared with the ACGAN on generating images using the CIFAR10 dataset [18]. The classification error is presented, comparing the two methods.

All the networks are trained in Lasagne [13] on top of Theano [14] in Python, unless stated otherwise.

Note that evaluating any GAN implementation is still an open problem since there is not any consistent measurement in evaluating a deep generator. As explained in [19], "As of yet, there is no consensus regarding the best score." In fact some scores like structural similarity index (SSIM) and Universal Quality Index (UQI) favor the models closer to a random generator. Other measurements such as Inception Score use another pre-trained deep neural networks (which is biased towards a specific database) to evaluate a deep generator. This is not a valid evaluation except if the generator is trained on the exact same database (this is presented in the second experiment with the CIFAR10 database [18]). So far the most accepted evaluation for deep generators is the visual correctness of the generated samples. Therefore, in the first experiment, the visual results are presented for the MNIST [17] dataset. For the second experiment which is performed on the CIFAR10 dataset [18], since the images are very small to represent visual information, the classification error and confusion matrices are presented to evaluate the deep generators.

1) MULTI-CLASS VAC+GAN WITH MNIST
a: DATASET
MNIST ("Modified National Institute of Standards and Technology") [17] is known as the "hello world" dataset of computer vision. It is a historically significant image classification benchmark introduced in 1999, and there has been a considerable amount of research published on MNIST image classification. MNIST contains 60,000 training images and 10,000 test images, both drawn from the same distribution. It consists of 28×28 pixel images of handwritten digits. Each image is assigned a single truth label digit from [0, 9].

b: NETWORK ARCHITECTURE
The proposed method of the multi-class VAC+GAN has been applied to the DCGAN structure [6]. The generator, discriminator and the classifier used in this experiment are given in Table 2, 3 and 4 respectively. All convolutions and deconvolution layers in Table 2, 3 are using (2,2) padding with stride (2,2).

c: IMPLEMENTATION VAC+GAN
The loss functions for the proposed VAC+GAN method for multi-class scenarios, (using MNIST) is given by:

$$L_g = \vartheta \cdot BCE(G(z|c), 1) + \zeta \cdot CCE$$
$$L_d = BCE(x, 1) + BCE(G(z|c), 0) \qquad (31)$$

where, L_g, and L_d are the generator and discriminator losses respectively, G is the generator function, BCE is the binary cross-entropy loss for discriminator and CCE is the categorical cross-entropy loss for the classifier. For this experiment the hyper-parameters of the loss function, ϑ and ζ are equal to 0.2 and 0.8 respectively.

The optimizer used for training the generator and discriminator is ADAM with learning rate, β_1 and β_2 equal

TABLE 2. The generator's structure for the experiment in IV-B1.

Layer	Type	kernel	Activation
Input	Input(10)	–	–
Hidden 1	Dense	1024	ReLU
BatchNorm 1	–	–	–
Hidden 2	Dense	$128 \times 7 \times 7$	ReLU
BathNorm 2	–	–	–
Hidden 3	Deconv	5×5 (64ch)	ReLU
BathNorm 3	–	–	–
Output	Deconv	5×5 (1ch)	Sigmoid

TABLE 3. The discriminator's structure for the experiment in IV-B1.

Layer	Type	kernel	Activation
Input	Input	–	–
Hidden 1	Conv	5×5 (64 ch)	LeakyR(0.2)
BatchNorm 1	–	–	–
Hidden 2	Conv	5×5 (128 ch)	LeakyR(0.2)
BathNorm 2	–	–	–
Hidden 3	Dense	1024	LeakyR(0.2)
Output	Dense	1	Sigmoid

TABLE 4. The classifier's structure for the experiment in IV-B1.

Layer	Type	Kernel	Activation
Input	Input(28×28)	–	–
Hidden 1	Conv	3×3(16 ch)	ReLU
Pool 1	Max pooling	2×2	–
Hidden 2	Conv	3×3(8 ch)	ReLU
Pool 2	Max pooling	2×2	–
Hidden 3	Dense	1024	ReLU
Output	Dense	10	Softmax

to 0.0002, 0.5 and 0.999 respectively. And the classifier is optimized using nestrov momentum gradient descent with learning rate and momentum equal to 0.01 and 0.9 respectively.

d: RESULTS

In Fig. 8a-8d the generated MNIST digits from 4 different methods are illustrated. The results from the CGAN [7] and CDCGAN [6] are shown in Fig. 8a and 8b. It is noticeable that the quality of the generated digits from the CGAN and CDCGAN is not sufficient and in many cases the output is vague. The proposed VAC+GAN method is able to generate digits of superior quality (Fig. 8d) compared to CGAN and CDCGAN, while using the exact same structure of generator as in CDCGAN. The results of ACGAN in Fig. 8c are comparable with the results generated from the VAC+GAN method, but the main advantage of the proposed method over the ACGAN is its versatility as it can be applied to any GAN implementation regardless of the model's architecture and loss function.

2) MULTI-CLASS VAC+GAN WITH CIFAR10
a: DATASET

In this experiment the CIFAR10 database [20] is utilized, which consists of 60000 images in 10 classes. In this work the existing default split was used, wherein 50000 of these images are for training and 10000 for testing purposes.

TABLE 5. The generator's structure for the experiment in IV-B2.

Layer	Type	kernel	Activation
Input	Input	–	–
Hidden 1	Dense	384×4	ReLU
Reshape	Reshape	384ch 4×4	–
Hidden 2	DeConv	5×5 (192 ch)	ReLU
BathNorm 2	–	–	–
Hidden 3	DeConv	5×5 (96 ch)	ReLU
BathNorm 3	–	–	–
Output	DeConv	5×5 (3 ch)	tanh

TABLE 6. The discriminator's structure for the experiment in IV-B2. The *st* stands for stride size and MBDisc for Mini Batch Discrimination layer.

| Layer	Type	kernel	Activation
Input	Input	$32 \times 32 \times 3$	–
Gaussian	Noise	$\sigma = 0.05$	–
Hidden 1	Conv	16ch $st(2, 2)$	LeakyR(0.2)
DropOut 1	DropOut	$p = 0.5$	–
Hidden 2	Conv	32ch $st(1, 1)$	LeakyR(0.2)
BathNorm 1	–	–	–
DropOut 2	DropOut	$p = 0.5$	–
Hidden 3	Conv	64ch $st(2, 2)$	LeakyR(0.2)
BathNorm 2	–	–	–
DropOut 3	DropOut	$p = 0.5$	–
Hidden 4	Conv	128ch $st(1, 1)$	LeakyR(0.2)
BathNorm 3	–	–	–
DropOut 4	DropOut	$p = 0.5$	–
Hidden 5	Conv	256ch $st(2, 2)$	LeakyR(0.2)
BathNorm 4	–	–	–
DropOut 5	DropOut	$p = 0.5$	–
Hidden 6	Conv	512ch $st(1, 1)$	LeakyR(0.2)
BathNorm 5	–	–	–
DropOut 6	DropOut	$p = 0.5$	–
MBDisc [21]	–	–	–
Output	Dense	1	sigmoid

b: NETWORKS ARCHITECTURE

For this experiment, the proposed VAC+GAN multi-class scheme is compared with the ACGAN [8]. The networks utilized in this experiment are shown in Table 5, 6 and 7 corresponding to the generator, discriminator[3] and classifier respectively. All deconvolution layers in Table 5 are using 'SAME' padding with stride (2,2). Also in Table 6, the deconvolution layers are using 'SAME' padding with kernel size 3×3. The same generator and discriminator architectures have been used in both VAC+GAN and ACGAN method [8] to obtain fair comparisons.

c: IMPLEMENTATION VAC+GAN

The loss functions for the proposed VAC+GAN method for multi-class scenarios, (using CIFAR) is the same as in the previous experiment (with MNIST) given in Equation 31 and used to train the models. In this experiment, ϑ and ζ are equal to 0.2 and 0.8 respectively.

The optimizer used for training the generator and discriminator is ADAM with learning rate, β_1 and β_2 equal to 0.0002, 0.5 and 0.999 respectively. The classifier is optimized using nestrov momentum gradient descent with learning rate and momentum equal to 0.01 and 0.9 respectively.
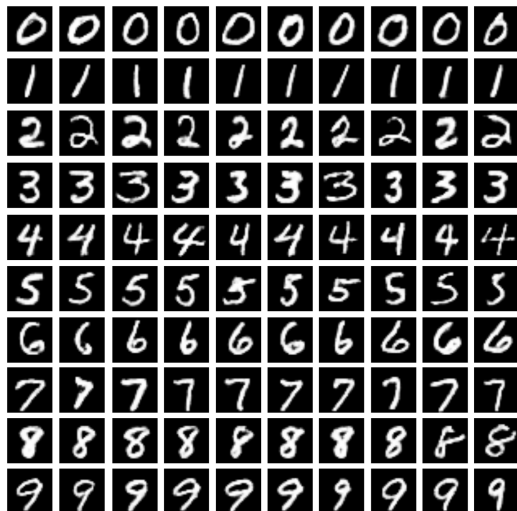

[3]https://github.com/King-Of-Knights/Keras-ACGAN-CIFAR10/blob/master/cifar10.py
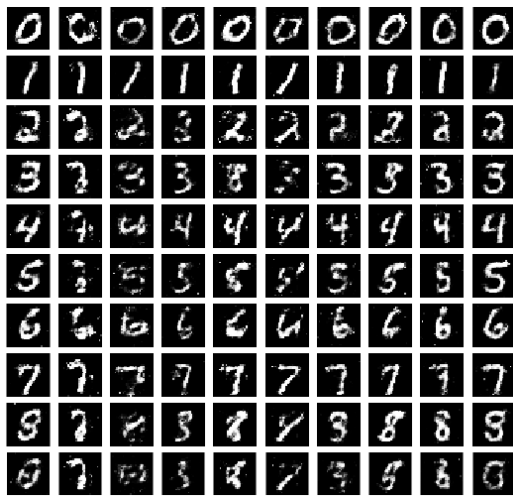
(a) Samples drawn from conditional generator trained using CGAN [7] on MNIST [17] dataset. Each row corresponds to one class.
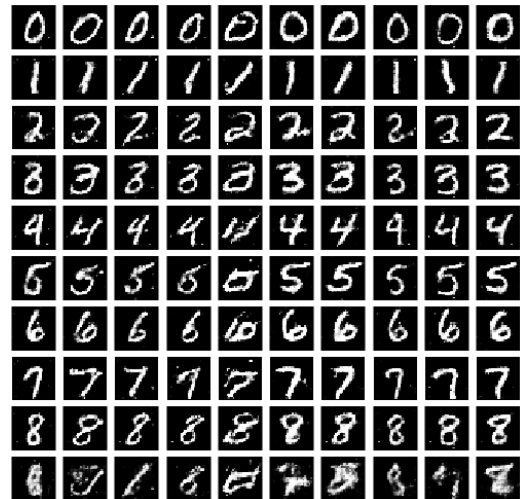
(b) Samples drawn from conditional generator trained using CD-CGAN [6] on MNIST dataset [17]. Each row corresponds to one class.

(c) Samples drawn from conditional generator trained using ACGAN [8] on MNIST dataset [17]. Each row corresponds to one class.
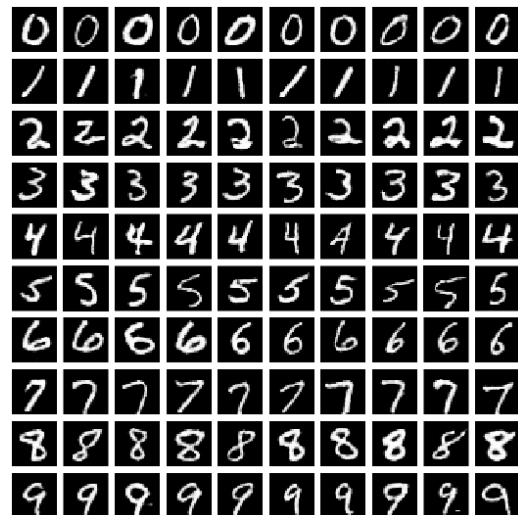
(d) Samples drawn from conditional generator trained using VAC+GAN on MNIST dataset [17]. Each row corresponds to one class.

FIGURE 8. Deep conditional generators trained on MNIST [17].

d: RESULTS

The results of the ACGAN [8] and the proposed method of VAC+GAN, using the CIFAR10 database [18], are shown in Fig. 9a[4] and 9b respectively. The CIFAR10 database is extremely unconstrained and there are just 1000 samples in each class. Consequently the outputs of both implementations are vague. Therefore, in order to compare the two methods, their classification errors are compared. The confusion matrix for ACGAN [8] and VAC+GAN are shown in Fig. 10a[5] and 10b respectively. The confusion matrices show that the

proposed VAC+GAN method has a better classification performance compared to the ACGAN. The classification accuracy of the ACGAN and the VAC+GAN on CIFAR10 is 71.89% and 74.49% respectively after 200 epochs. The proposed method not only has a higher classification accuracy but also has the main advantage of versatility. In VAC+GAN scheme the most appropriate classification network can be selected in contrast with the ACGAN method where the classification task is restrained to the discriminator which performs as the classifier as well. Moreover the VAC+GAN can be applied to any GAN implementation just by placing a classifier in parallel with discriminator as shown through the mathematical proofs given in III.

[4]https://github.com/King-Of-Knights/Keras-ACGAN-CIFAR10/blob/master/plot_epoch_220_generated.png

[5]https://github.com/King-Of-Knights/Keras-ACGAN-CIFAR10/blob/master/Confusion_Matrix.png

TABLE 7. The classifier's structure for the experiment in IV-B2.

| Layer	Type	kernel	Activation
Input	Input	$32 \times 32 \times 3$	–
Hidden 1	Conv	5×5 (128ch)	ReLU
BatchNorm 1	–	–	–
MaxPool 1	MaxPool	(2,2)	–
Hidden 2	Conv	5×5 (256ch)	ReLU
BatchNorm 2	–	–	–
MaxPool 2	MaxPool	(2,2)	–
Hidden 3	Conv	5×5 (512ch)	ReLU
BatchNorm 3	–	–	–
MaxPool 3	MaxPool	(2,2)	–
Hidden 4	Dense	512	ReLU
Output	Dense	10	softmax

C. VAR+GAN

In this section, the implementation of the VAR+GAN is presented and compared to the cBiGAN method [9] in the task of generating face samples in a particular set of landmark points using the CelebA [12] dataset. To keep the consistency in comparisons, the same architecture for the generator network has been kept throughout all implementations. All the networks are trained using the Lasagne [13] library on the top of Theano [14] in Python.

1) DATABASE

The dataset used in this experiment is the CelebA database [12] which is made of 202,599 frontal posed images. Similarly to the experiments in the binary VAC+GAN, the face regions are cropped and resized to 48×48 pixels using OpenCV frontal face cascade classifier [22]. Supervised Descent Method (SDM) [23] is used for facial landmark points detection. The detector is based on [24] and it utilizes the discriminative 3D facial deformable model to find 49 facial landmarks including contours of eyebrows, eyes, mouth and the nose. These landmarks are used as the data aspect in this experiment.

2) NETWORK ARCHITECTURES

The three architectures used in this experiment are the encoder, the decoder, and the regression networks. The first two are shown in Fig. 3. The encoder network is made of convolutional layers with ELU activation function. The downscaling in these layers is obtained by using (2, 2) stride in every second convolutional layer. In the decoder network, all convolutional layers have 64 channels while in the encoder, the number of the channels is gradually increased to 128, 192, and 256 after each pooling layer. The decoder contains one fully connected layer which reshapes the input to match the dimensionality of the next layer. Next layers are all convolutional layers followed by (2, 2) un-pooling layers for every second convolution. The exponential linear unit (ELU) [16] is used as activation function except in the last layer wherein no non-linearity has been applied except for the encoder in cBiGAN scheme wherein tanh nonlinearity is applied in the output layer. For the layers shown in red no non-linearity is applied to the input. The regression network is shown in table 8.
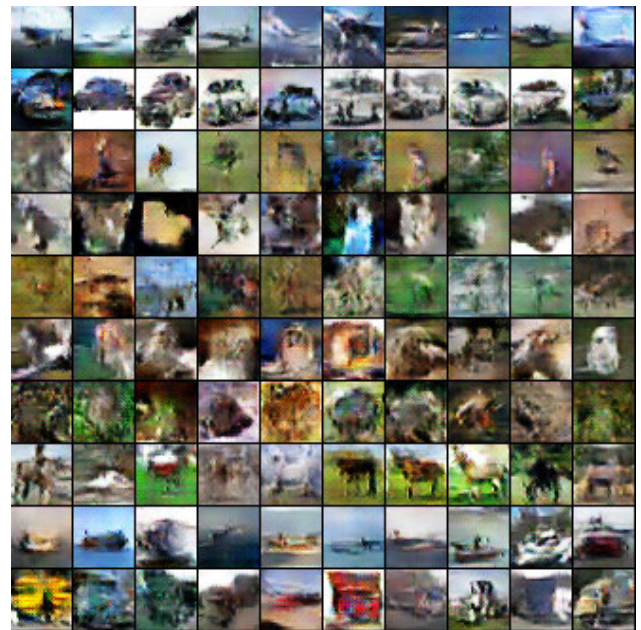
(a) Generated samples of ACGAN [8]
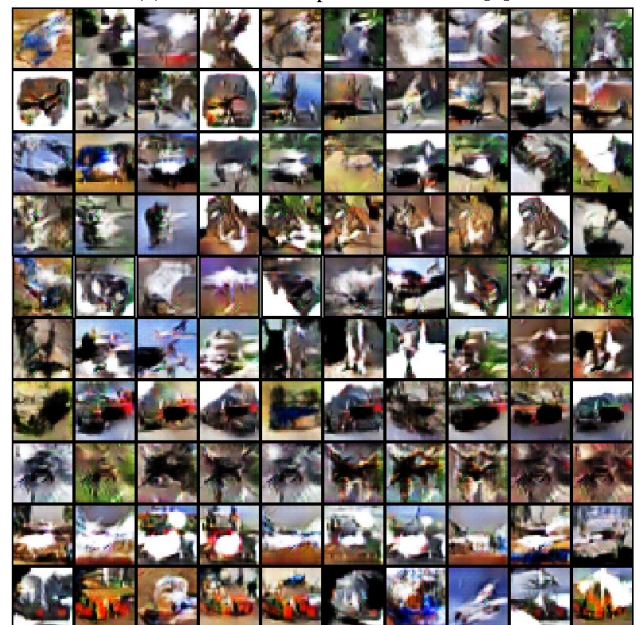

(b) Generated samples of VAC+GAN.

FIGURE 9. Generated samples using ACGAN [8] and the proposed method VAR+GAN method on CIFAR10 [18]. Each row corresponds to a class.

3) IMPLEMENTATION

a: VAR+GAN

For the proposed scheme (VAR+GAN) (Fig. 2b), the Boundary Equilibrium Generative Adversarial Network (BEGAN) [5] is utilized to train the deep generator. In this method the generator's and discriminator's input dimension is $k = 128$. The loss function for the proposed implementation is a modified version of the original BEGAN loss [5] given by:

$$L_d = L(x) - k_t \cdot L(G(z|y))$$
$$L_g = \vartheta \cdot L(G(z|y)) + \zeta \cdot L_R$$
$$k_{t+1} = k_t + \lambda_k \big(\gamma L(x) - L(G(z|y)) \big) \tag{32}$$
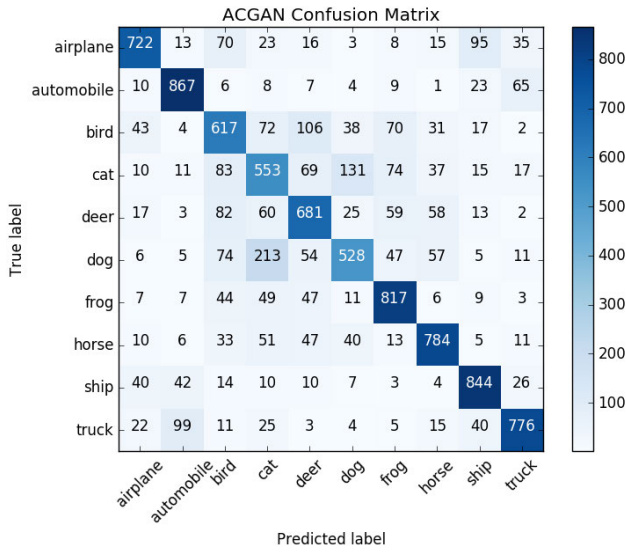
(a) Confusion matrix for ACGAN [8] method on CIFAR10 [18]
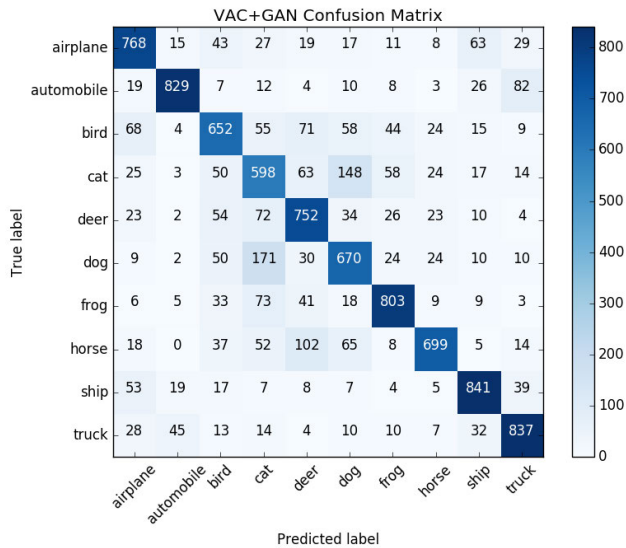
(b) Confusion matrix for VAC+GAN method on CIFAR10 [18]

**FIGURE 10.** Confusion matrices for the classifier trained on ACGAN [8] and VAC+GAN methods on CIFAR10 [18].

TABLE 8. The regression network used in the experiments described in IV-B2.

Layer	Type	kernel	Activation
Input	Input(48×48)	–	–
Hidden 1	Conv	3×3(64 ch)	ReLU
Pool 1	Max pooling	2×2	–
Hidden 2	Conv	3×3(64 ch)	ReLU
Pool 2	Max pooling	2×2	–
Hidden 3	Dense	1024	ReLU
Output	Dense	98	Tanh

where L_g and L_d are generators and discriminators losses respectively. G is the generator function, z is a sample from the latent space, x and y are genuine image and corresponding ground truth drawn from the database, λ_k is the learning rate for k, γ is the equilibrium hyper parameter set to 0.5 in this work, L_R is the regression loss given by equation 14, and

(a) Proposed VAR+GAN method on CelebA [12]
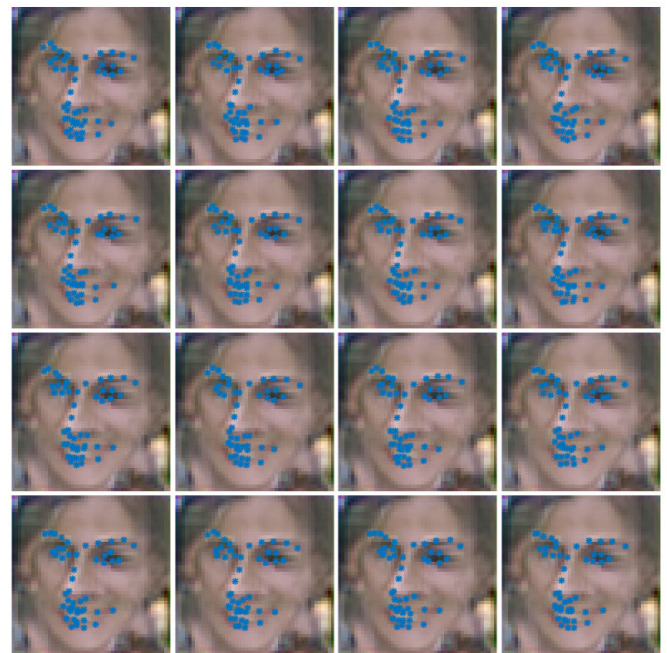
(b) cBiGAN [9] on CelebA [12]

FIGURE 11. Generator outputs for proposed VAR+GAN method vs cBiGAN [9] on the CelebA [12]. (Experiment described in IV-C).

ϑ and ζ are set to 0.97 and 0.03 respectively, and L is the auto-encoders loss defined by

$$L(v) = |v - D(v)|^2 \qquad (33)$$

The optimizer used for training the generator and discriminator is ADAM with learning rate, β_1 and β_2 equal to 0.0001, 0.5 and 0.999 respectively. And the regression network is optimized using nestrov momentum gradient descent with learning rate and momentum equal to 0.01 and 0.9 respectively.

(a) Proposed VAR+GAN method.

(b) cBiGAN [9]

FIGURE 12. Generator outputs for proposed VAR+GAN method vs cBiGAN [9] on the CelebA [12], given particular landmarks. (Experiment described in IV-C).

(a) Proposed VAR+GAN method.

(b) cBiGAN [9]

FIGURE 13. Generator outputs for proposed VAR+GAN method vs cBiGAN [9] on the CelebA [12], given particular landmarks. (Experiment described in IV-C).

b: cBiGAN

The cBiGAN scheme [9] is implemented in the following way. The generator's architecture is same as the decoder network shown in Fig. 3b with input dimension $k = 128$. The discriminator model is same as the encoder network in Fig. 3a with $k = 1$ and *sigmoid* non-linearity at the output layer. And the encoder network in Fig. 1b has the architecture shown

in Fig. 3a with $k = 98$ and *tanh* non-linearity at the output layer. The loss function for this scheme is presented in [9] given by

$$L_D = \log(p^r) + \log(1 - p^I) + \log(1 - p^s)$$
$$L_G = \log(p^I)$$
$$L_E = \log(p^s) + \theta ||s^- - y||^2 \qquad (34)$$

wherein s^- is the encoder's output, y is the genuine aspect coming from the database, and

$$p^r = D(y, x) \quad , \quad p^I = D(y, G(z|y)) \quad , \quad p^s = D(s^-, x)$$
(35)

where D and G are discriminator and generator functions respectively, x and y are genuine image and corresponding ground truth drawn from the database, and z is a sample from the latent space. The coefficient θ is set to 0.8. The optimizer used for training the model is ADAM with learning rate, β_1 and β_2 equal to 0.0001, 0.5 and 0.999 respectively.

4) RESULTS

In this section, the proposed VAR+GAN method is compared against the cBiGAN method [9] while generating faces for a particular landmark point set. The results for six sets of landmarks are shown in Fig. 11 to 13. In each figure the outputs from the proposed method for a particular set of landmarks are illustrated in 11a,12a and 13a while in the Fig. 11b, 12b and 13b, the output of the generator trained in cBiGAN scheme [9] is given for the same landmarks.

As shown in Fig. 11 - 13, both methods are able to generate samples constrained to a particular set of landmarks but the proposed method generates higher variations of faces for a given landmark set while cBiGAN [9] fails to create different samples in the same condition. Also the advantage of VAR+GAN lies is the versatility of the method which facilitates the implementation and also guarantees the higher quality in the generated samples. As prime example, in this work the proposed VAR+GAN method is taking advantage of the simplicity and power of BEGAN implementation [5] and the only change applied, is to place the regression network and add its error value to the generator's loss, while cBiGAN method [9] is constrained to a specific loss function which remains a disadvantage.

V. DISCUSSION AND CONCLUSION

The main contributions of this work is the introduction of two new methods to train conditional deep generators along with their mathematical proofs. The first method is called Versatile Auxiliary Classifier with Generative Adversarial Network (VAC+GAN) which can be used for both binary and multi-class-classification problems. The second approach is an extension of the initial idea of VAC+GAN, used in regression tasks. It is called Versatile Auxiliary Regression with Generative Adversarial Network (VAR+GAN). In the VAR+GAN scheme a new loss function is also proposed. The main idea is to place a classification/ regression network in parallel to the discriminator network and back-propagate the classification/ regression loss through the generator network in the training stage. The mathematical proofs provided show that the proposed methods are versatile to be applicable to any GAN structure with any loss function, as well as having the advantage of choosing any architecture for the classification/ regression network. For both schemes, the mathematical proofs show that the presented frameworks increase the JSD between samples/classes generated by the deep generator i.e.

the generator can produce more distinct samples, which is desirable. The code for both methods is available at [1] and [2].

The proposed schemes are implemented and compared with similar state-of-art methods. Firstly, the proposed VAC+GAN is presented for both binary and multi-class classification. The VAC+GAN for the binary classification case is implemented for training a gender specified generator. The results has been compared to another versatile method known as Conditional GAN (CGAN) [7], showing the superiority of the proposed VAC+GAN method. Furthermore the VAC+GAN is implemented for multi-class cases and the results have been compared to the implementation of CGAN [7], CDCGAN [6] and ACGAN [8] on the MNIST dataset [17]. Also experiments conducted on the CIFAR10 dataset [18] and comparisons are made with respect to the ACGAN method [8]. In all cases the VAC+GAN has given superior results compared to the other methods.

Finally, the proposed VAR+GAN method is compared with a state-of-the-art method with the same purpose. The cBiGAN method [9] generates samples with a particular aspect the variations between the generated samples is limited, while the proposed VAR+GAN produces greater variations for a specific set of aspects. Being able to generate variable samples is crucial for tasks including augmentation purposes.

Overall the proposed methods of VAC+GAN and VAR+GAN show superior results when compared with similar state-of-the-art techniques. Except the increase in the performance for several tasks, the main advantage is their versatility, as the proposed schemes can be applied to any GAN structure with any loss function as well as having the advantage of choosing any architecture for the classification/regression network.

The future work includes applying the method to datasets with a larger number of classes (such as CIFAR-100 [18], ImageNet 1000 [25]) for more complex tasks, extend the implementation for images of higher resolution and experiment with various GAN architectures. Also comparisons with more SoA methods will be conducted to study extensively the proposed method. Additional, future studies involve merging the VAC+GAN and VAR+GAN methods to constrain the generator to create samples from a specific class with a particular continuous aspect and also investigating the influence of the generated samples in augmentation task for different applications.

APPENDIX
DIVERSITY MEASUREMENTS

1) **MSE (Mean Squared Error)**: MSE measures the average of the squares of the errors or deviations; representing the difference between the estimator and what is estimated. The lower value of MSE shows lesser error.

$$MSE(f, g) = \frac{1}{mn} \sum_{0}^{m-1} \sum_{0}^{n-1} ||f(i,j) - g(i,j)|| \quad (36)$$

2) **RMSE (Root Mean Squared Error)**: RMSE is a quadratic scoring rule that measures the average magnitude of the error. It is the square root of the average of squared differences between prediction and actual observation. The lower value of RMSE shows lesser error.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (37)$$

3) **MAE (Mean Absolute Error)**: MAE also measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The lower value of MAE shows lesser error.

$$MAE(f, y) = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| \qquad (38)$$

4) **UQI (Universal Quality Index)** [26]: UQI measures the structural distortion of the images by modeling the distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. The higher value of UQI shows lesser error.

5) **SSIM (Structural Similarity Index)** [27]: SSIM is a perception-based model that considers image degradation as perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms. The higher value of SSIM shows lesser error.
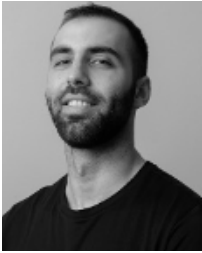
The MSE, RMSE and MAE show the difference between two images. Higher values for these metrics correspond to higher variation of the generated images. UQI and SSIM measure the structural similarity between two samples. Lower value for these metrics corresponds to reduced similarity. Therefore when evaluating generative models, higher values for MSE, RMSE, and MAE and lower values for UQI and SSIM are desirable.

REFERENCES

[1] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision," *IEEE Consum. Electron. Mag.*, vol. 6, no. 2, pp. 48–56, Apr. 2017.

[2] 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.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[4] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*. [Online]. Available: http://arxiv.org/abs/1609.03126

[5] 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

[6] 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

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[8] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," 2016, *arXiv:1610.09585*. [Online]. Available: http://arxiv.org/abs/1610.09585

[9] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.

[10] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," 2016, *arXiv:1606.03657*. [Online]. Available: http://arxiv.org/abs/1606.03657

[11] X. Li, L. Chen, L. Wang, P. Wu, and W. Tong, "SCGAN: Disentangled representation learning by adding similarity constraint on generative adversarial nets," *IEEE Access*, vol. 7, pp. 147928–147938, 2019.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[13] S. Dieleman, *Lasagne: First Release, Version v0.1*. Zenodo, Aug. 2015. [Online]. Available: https://zenodo.org, doi: 10.5281/zenodo.27878.

[14] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, D. Warde-Farley, I. Goodfellow, A. Bergeron, and Y. Bengio, "Theano: Deep learning on GPUs with Python," in *Proc. NIPS, BigLearning Workshop*, Granada, Spain, vol. 3, 2011, pp. 1–48.

[15] *OpenCV Face Detection Using Haar Cascades*, Open Source Comput. Vis.-OpenCV, 2001. [Online]. Available: https://opencv.org

[16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289

[17] Y. LeCun. (1998). *The MNIST Database of Handwritten Digits*. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[18] A. Krizhevsky, V. Nair, and G. Hinton. (2009). *CIFAR-10 and CIFAR-100 Datasets*. [Online]. Available: https://www.cs.toronto.edu/kriz/cifar.html

[19] A. Borji, "Pros and cons of GAN evaluation measures," 2018, *arXiv:1802.03446*. [Online]. Available: http://arxiv.org/abs/1802.03446

[20] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.

[21] 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.

[22] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2006, p. 3.

[23] X. Xiong and F. D. L. Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[24] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1859–1866.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

SHABAB BAZRAFKAN (Student Member, IEEE) received the B.Sc. degree in electrical engineering from Urmia University, Urmia, Iran, in 2011, the M.Sc. degree in telecommunication engineering, image processing branch from the Shiraz University of Technology (SuTECH), in 2013, and the Ph.D. degree in deep learning and neural network design from the National University of Ireland Galway (NUIG), in 2018. He is currently a Postdoctoral Researcher working on low dose CT image reconstruction using machine learning techniques with the Vision Laboratory, University of Antwerp.

VIKTOR VARKARAKIS (Graduate Student Member, IEEE) received the B.Sc. degree in computer science and intelligent systems from the University of Piraeus, Greece, in 2017. He is currently pursuing the Ph.D. degree with the National University of Ireland, Galway (NUIG). He is also with FotoNation/Xperi. His research interests include machine learning using deep neural networks for tasks related to computer vision.

HOSSEIN JAVIDNIA received the master's degree in computer network security from the University of Guilan, in 2015, and the Ph.D. degree in electronic and electrical engineering from the National University of Ireland Galway, in 2018, examining the "Contributions to the Measurement of Depth in Consumer Imaging." After graduation, he started working as an Research and Development Engineer at Xperi Corporation, mainly focused on consumer 3D reconstruction. He was also the Lead Researcher on a collaborative project with Huawei Ltd., Ireland "Deep Real: Deep Learning for 3D Object placement in videos" nominated for Technology Award Ireland 2019. He is currently a Research Fellow with the ADAPT Centre, Trinity College Dublin, focused on advanced 3D modeling and augmented reality. His current research interests include 3D reconstruction from large and small motions, simultaneous localization and mapping, and application of artificial intelligence in computer vision tasks.

JOSEPH LEMLEY (Member, IEEE) received the B.S. degree in computer science and the M.S. degree in computational science from Central Washington University, Ellensburg, WA, USA, in 2006 and 2016, respectively, and the Ph.D. degree from the National University of Ireland, Galway. He is currently leading Xperi's Sensing Group, which develops novel algorithms and artificial neural networks for upcoming sensor technologies. His research interests include artificial intelligence, deep learning, and computer vision. He received the 2017 Best Paper Joint Award for the *IEEE Consumer Electronics Magazine*, and the Best Paper Second Place Award at ICCE 2018 and other awards during previous years.

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 the 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*.

● ● ●

Appendix I

Deep Learning for Consumer Devices and Services 2 – AI Gets Embedded at the Edge

Cover Article: Edge-AI

Deep Learning for Consumer Devices and Services 2—AI Gets Embedded at the Edge

**Peter Corcoran, Joseph Lemley,
Claudia Costache, and Viktor Varkarakis**
National University of Ireland Galway (NUIG)

Abstract—The recent explosive growth of deep learning is enabling a new generation of intelligent consumer devices. Specialized deep learning inference now provides data analysis capabilities that once required an active cloud connection, while reducing latency and enhancing data privacy. This paper addresses current progress in Edge artificial intelligence (AI) technology in several consumer contexts including privacy, biometrics, eye gaze, driver monitoring systems, and more. New developments and challenges in edge hardware and emerging opportunities are identified. Our previous article, "Deep learning for consumer devices and services," introduced many of the basics of deep learning and AI. In this paper, we explore the current paradigm shift of AI from the data center into CE devices—"Edge-AI."

■ THIS PAPER FOLLOWS the earlier publication,[1] "Deep learning for consumer devices and services: Pushing the limits for machine learning, and computer vision." That article introduced the basics of deep learning along with the supporting tools and methodologies. Our vision at

that time was that new embedded hardware solutions would enable advanced capabilities and features incorporating convolutional neural network (CNN)-based AI across a broad range of Consumer Electronic (CE) devices and services.

Since that time, there has been a growing interest and investment by industry into moving key elements of AI away from the cloud toward the sensors and the embedded devices themselves. The movement of AI closer to the device was

covered by two special issues of CE magazine[2,3] near the end of 2016. At that time, industry was focused on OpenFog, an initiative to define a new generation of low-latency services for the Internet-of-things (IoT). But, AI is now moving onto the device itself with many companies and researchers focusing on developing FPGA-based solutions[4] and, most recently, embedded AI hardware accelerators.[5,6]

Most of the large semiconductor manufacturers are working on a new generation of AI accelerator chipsets and the widespread deployment of neural networks on the device itself. In fact, such technology is already incorporated into the latest generation of mobile handsets, high-end television panels, professional digital cameras, and many new automotive subsystems. This embedding of AI into the device itself is referred to as *Edge-AI* and is distinguished from AI services provided over a low-latency network link, better known as *Mobile-Edge AI*.

There is a role for both network-based and device-based AI, but it is the recent emergence of *on-device* implementations that is most exciting for CE engineers. Improvements in the computational and energy efficiencies of hardware AI accelerators over today's GPU-based solutions provides an ideal solution for challenging data-processing problems introduced by new battery-powered, wearable, and IoT devices.

In this paper, progress in *Edge-AI* over the past two years is reviewed and several examples of practical problems tackled with CNNs are outlined. Some are contributions made within our research team, others are drawn from the literature, but each example illustrates how CNN-based *Edge-AI* will be at the core of many new devices, systems, and services that emerge over the next decade.

EXAMPLE CE USE CASES AND DL SOLUTIONS

A good starting point is to consider how *Edge-AI* solutions can improve performance and operational efficiency to a point where the benefits outweigh the costs of incorporating an inference engine or platform into the system. Looking at the research literature and the evolution of consumer electronics products over the past decade, one area where DL can add value is in computer vision applications. The cost of a complete VGA camera module has dropped below $1.00 and the cost of adding a CMOS image sensor is almost negligible in today's devices. Thus, many interesting uses of *Edge-AI* focus on combining advanced image analysis capabilities with low-cost CMOS sensors. Following are some good examples of the associated CE applications that *Edge-AI* is enabling today.

Eye-Gaze Systems

A range of applications based on eye-gaze tracking in consumer platforms such as automotive (for driver monitoring), augmented and virtual reality (AR/VR) (for foveated rendering and immersive experiences), and smartphones and TV (for gaze-based user interfaces and saliency analysis of content) have been described in a recent review.[7] Recent research in this field includes the work presented by Lohmeyer et al.[8] where gaze duration and patterns are used to assess how effectively users can operate a connected self-injection system. Similarly, the effectiveness of observation charts in a hospital is evaluated[9] by comparing user viewing patterns derived from eye tracking data. Eye movements and pupillary response are used as indicators of cognitive load.[10,11] In the work of Przybyło et al.,[12] the influence of emotions on the visual acuity of users was studied, which showed that eye movements like fixations and saccades clearly respond to levels of stress. Eye gaze estimation is a "must have" feature for the latest driver monitoring systems (DMS) such as illustrated in Figure 1 and is crucial for the functioning of AR/VR systems. For these applications, deep learning can either be applied as an end-to-end solution[13] or as a component of a more traditional gaze estimation pipeline, such as iris segmentation.[14,15]

End-to-end approaches to eye gaze estimation are state of the art for uncontrolled environments where the camera is at a distance, and such methods are closing the gap in AR/VR settings.[16] Real-world *Edge-AI* must balance accuracy requirements with power availability and speed. Typical Edge products have strict limits on the number of MAC operations per second, and these limits often preclude the use of the largest and most popular networks without

Figure 1. State-of-the-art DMS—screenshots courtesy of Xperi, Inc.

Figure 2. User-facing cameras for a mixed reality headset—courtesy of pupil-labs.com.[35]

significant optimization. Similar approaches can have vastly different performance. Many papers report performance on GPUs, but it is important to consider their speed on the low-cost commodity processors that are typical of Edge devices.

For example, one published gaze estimation solution operates at approximately 120 fps on a GPU, but 0.0043 fps on an ARM processor (~one frame every 4 min), while another network with similar accuracy runs at 92.59 fps on the same ARM processor, or 19 fps with 1.5% higher accuracy.[13]

Biometrics and Device Authentication

Biometric authentication has the potential to solve many pressing problems related to security and privacy for today's CE devices.[17] This can be seen in today's mobile devices, but the real breakthrough will come when biometric technology can be incorporated into lower power devices such as wearables and IoT peripherals. A key aspect here is that biometric acquisition and processing should occur on the device or peripheral to secure the biometric.[18–20] Realistically, this can only be achieved by leveraging *Edge-AI* to enable secure, energy-efficient authentication of the biometric.

Automotive biometrics offer a compelling use case for *Edge-AI*. Our cars already have cameras to monitor the driver so it is straightforward to incorporate technology to verify driver identity. External cameras, for pedestrian detection can be repurposed for an initial authentication, supplementing today's keyless entry systems. In this use case, the key challenge is acquiring a biometric in unconstrained conditions.[21,22] While facial biometrics offer a good level of verifiability, they are prone to attacks[23–25] and even iris biometrics can be challenged.[26,27]

Incorporating all the elements of a biometric authentication chain—unconstrained acquisition, liveness verification, and, finally, authentication of the registered biometric—requires significant computational resources and energy. Each of these elements can be tackled independently, and in parallel, by neural network-based solutions. As examples, large pose three-dimensional (3-D) facial segmentation,[28] facial segmentation and alignment[29], and iris segmentation[14] can all be tackled with CNNs. Liveness detection is also receiving much attention from researchers.[30–33]

Biometrics is also important for wearable devices such as the one shown in Figure 2. Consider the emerging category of AR and mixed-reality (MR) headsets where there is no keyboard. These devices must be connected to an online account. Fortunately, these AR/MR headsets must track eye-gaze to accurately render objects onto the real-world scene. To track eye-gaze, it requires a user-facing camera which monitors and segments iris and pupil regions (Figure 3) to determine eye-gaze. It is a straightforward task to add authentication of the segmented iris region.[34]

Immersive Audio for MR Headsets

MR implies combining a mixture of augmented visual elements with additional nonvisual sensory elements. In practice, audio augmentations can provide more compelling immersion than visual cues. However, maintaining the illusion of the real world when adding audio to the perceived environment is

challenging. Each individual has a unique ear canal and our brains process and perceive audio in a highly personalized way. And our audio senses are attuned to visual cues, subconsciously anticipating changes in environmental acoustics. Thus, in a large cathedral, you expect echoes from your footsteps and voice; in a room with carpets and soft furnishings, the acoustics should be subdued. Thus, acoustic cues should adapt to the surroundings of an MR-headset, or the illusion of immersion is lost.

But energy usage on a headset is more critical than on a smartphone, providing another excellent use-case for *Edge-AI*: scene analysis[36,37] and materials recognition[38] combined with depth[39,40] can help build a detailed analysis of the surrounding acoustic environment. And with state-of-the-art *Edge-AI*, multiple neural networks can operate in parallel at a fraction of the power budget for a GPU.

Image Signal Processing Pipeline in a Camera

Processing the raw Bayer data from an image sensor is a classic example where camera engineers and photographic experts have devoted many years of effort to create a specialized image processing pipeline (IPP), illustrated in Figure 4. Bryce Bayer's original patent[41] filed in 1976 is a classic example of an engineering compromise that works so well that it has become the basis of modern digital imaging. Until recently, this "magic" happened via complex set of image analysis and processing algorithms, the IPP.[42]

But now, thanks to the magic of deep learning methodologies, it has become feasible to replace the IPP in an imaging system. This work began with the demosaicking step of image conversion,[43,44] followed by the idea to replace the key steps of denoising and demosaicking from the IPP with a single CNN network.[45–47] Other authors have progressed to complete replacement of the IPP,[48] or alternatively, to learn the detailed camera model embodied in an existing IPP,[49,50] which can lead to reprogrammable IPPs. Imagine that you can completely reprogram how your camera captures and develops raw images "on the fly." Well, this approach is beginning to make its way into actual products, so expect to see some exciting new features in higher end digital cameras shortly!

Once you replace the IPP with a CNN, new ideas emerge such as "Learning to see in the
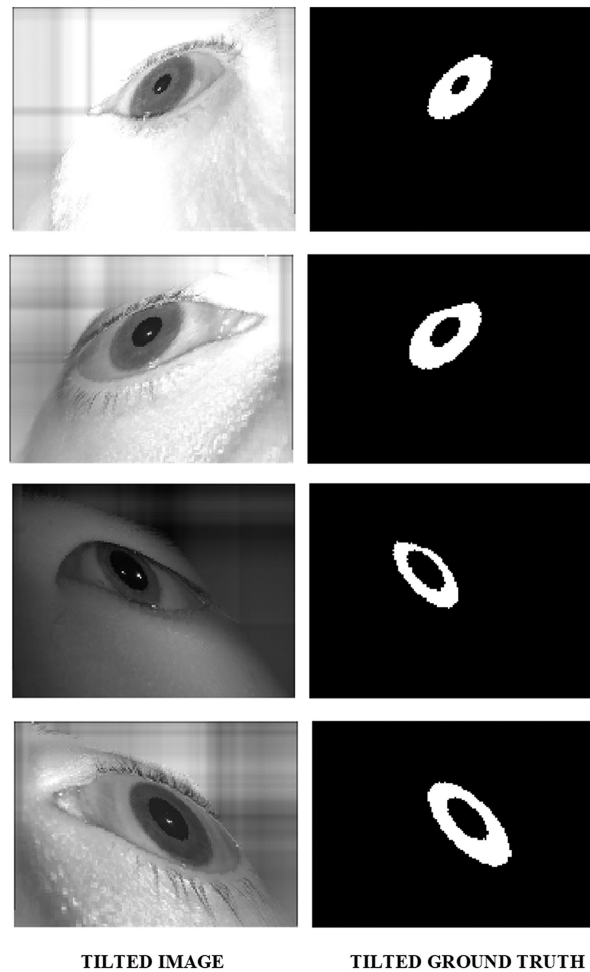

TILTED IMAGE **TILTED GROUND TRUTH**

Figure 3. Off-axis iris regions; accurate per-pixel segmentation is essential for user authentication.[15]

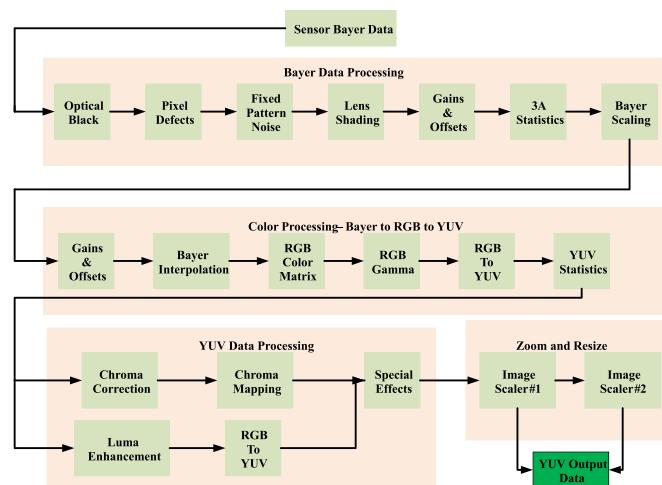dark."[51] In this particular example, researchers have used a CNN to solve a key challenge for

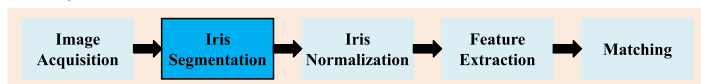**Figure 4.** Detailed view inside the IPP of a typical digital camera.[42]

Figure 5. Iris authentication processing pipeline.[15]

today's smartphone cameras—to capture images in low-illumination conditions.

CHALLENGES FOR AI DEPLOYMENTS IN CONSUMER ELECTRONICS?

Since our last article, there has been rapid progress with AI development, but many challenges specific to Edge-AI in consumer devices remain. Our recent work on a number of example CE-device problems has highlighted these challenges.

Problem-Specific Nature of AI Solutions

Every practical problem that we solve with *Edge-AI* is a component of a larger problem set. Focusing on a specific problem typically allows the research engineer to accurately define the data characteristics and the criteria to solve that particular problem. Let us consider the task of iris authentication where there is a processing pipeline, as shown in Figure 5.

Note that, in this sequence, some tasks can employ proven techniques such as iris-matching, deployed for more than two decades by the biometrics research community. However, the adoption of authentication on mobile and wearable devices introduces new tasks such as the segmentation and normalization of unconstrained iris regions to serve as input to established feature extraction and matching algorithms. These new challenges are crucial, and it is now well appreciated that iris segmentation is the predominant source of authentication error in mobile and wearable devices.[14,21]

Now, while deep learning is a powerful tool and can often achieve very impressive levels of accuracy, it is important to appreciate that neural networks can easily learn the wrong features from a poorly designed dataset and are vulnerable to adversarial attacks. This is a challenge we will return to later, and it remains an open-ended challenge.

Device-Specific Aspects

A unique aspect of applying deep learning techniques to consumer electronics is the device-specific nature of consumer data. Research has shown that images can be uniquely associated with a particular IPP.[49,50] This observation applies across other sensing capabilities of consumer devices. Thus, the collection of data from any particular consumer device, be it video, audio, six-axis motion or other forms of data collection, invariably exhibits unique characteristics.

Thus, to achieve optimal performance from a deep learning solution, the network should be trained on device-specific datasets. This is well known in the CE industry, where more traditional algorithms are tailored to individual models of production devices. In fact, variations in the manufacturing process, the local environment, or the calibration procedures applied can lead to differences in production batches of the same device. Thus, it is important to bear in mind that optimal performance of *Edge-AI* is achieved by tuning on device-specific data.

Conversely, a network that is tuned to a specific device may not perform well on other devices. While we have not explored this phenomenon across enough different problem cases, current experience suggests that a two-step approach makes sense. At stage one, a network is tuned on a generic dataset, representative of a range of similar data sensing systems (e.g., data acquired from multiple f2.0, 12 MP cameras). After this network is tuned to an acceptable level, then stage two involves additional tuning of some network layers on data from a specific production stream or batch of camera modules. Our experience shows that performance enhancements can be achieved, but the fine-tuned network cannot be used for other camera batches due to a loss of generality.

WHAT IS NEXT FOR *EDGE-AI*?

When we last considered the state-of-the-art, the AI-stick from *Movidius* was discussed as a practical example. Since that time, this company was acquired by Intel and a second generation of the AI-stick with enhanced performance is now available. There are now several other AI accelerators from mainstream players such as Nvidia's Jetson family of devices, and in mid-

2018 Google introduced an "Edge" version of its tensor processing unit. Outside of these mainstream players, there are many start-ups, spin-outs, and in-house projects working to deliver new low-power neural network accelerators.

Very soon, many of our readers will be developing systems and products based on these new AI platforms. *Edge-AI* with the promise of intelligent devices that have minimal power requirements—some able to run on a coin-sized battery for months without replacement—allows new CE devices to have functionality that only last year would require a large, power-hungry GPU or an always-on cloud connection. Our group has had the opportunity to build some interesting prototypes with these hardware accelerators, and some of you may have attended our "hands-on" workshop at IEEE GEM 2018, which was a great success; you can view a Twitter "moment" of the conference.[52] In our current graduate program lab classes, we help students build a handheld computer-vision terminal that can implement the Yolo one-shot object detector. This runs happily at 30 fps on a Raspberry PI coupled with the Intel AI-stick.

The age of embedded AI is now a practical reality. The open questions are how it will impact today's technology and what new challenges and opportunities the broader adoption of *Edge-AI* will bring.

Emerging Opportunities for Edge-AI

One area where *Edge-AI* will have enormous impact is on personal privacy. It is difficult to address privacy concerns when video data are constantly uploaded and processed in the cloud. By processing data on the device, one can avoid sending raw data over networks where it offers an attractive target for cyber criminals.

One example where *Edge-AI* will have a significant impact is in DMS, which will be mandated by the EU in 2022. These are already deployed in many high-end vehicles, and require devices that can instantly and intelligently react when a driver is distracted or impaired. They offer a stepping stone to fully autonomous vehicles, but pose a significant design challenge in the context of the EU's *general data protection regulation* (GDPR). While 5G technology can arguably perform the advanced processing required

by DMS, this approach requires moving data off-vehicle with associated data security and privacy issues. In contrast, *Edge-AI* enables onboard data processing and can employ a secure compute-unit within the sensing subsystem. Placing computation as close as possible to the sensor reduces latency and costs while addressing privacy of data.

Other opportunities lie in new wearable devices and smart cities. A new generation of smart glasses (e.g., Magic Leap and Hollolens) and wearable audio enhancement devices, known as hearables, represent devices that can directly modify our perception of the surrounding environment. The computational requirements to achieve real-time perceptual analysis followed by a realistic blending of additional visual and acoustic elements into the user experience are beyond the capabilities of today's embedded-GPU solutions. The answer lies with the next generation of *Edge-AI* hardware accelerators that can achieve the required real-time data processing rates with levels of energy efficiency orders of magnitude lower than is possible today.

In smart cities, we have an urban environment permeated with ubiquitous networks of sensors and services, but this poses significant challenges. How can we authenticate individuals in such an environment to validate their access to services and, more importantly, how do we guard the privacy of individuals when their every move is tracked by a multitude of cameras and sensing technologies? Again, *Edge-AI* can offer new solutions. Biometric processing can be implemented within devices so that registered users can be authenticated without a global sharing of their biometric data.[53] Once we have authenticated individuals, they can be linked with a global-ID that is independent of the local device authentication using techniques such as blockchain or zero-knowledge proof.[54] Then, once the individual is globally authenticated, they can be flagged with a "do not track" marker and compliance with regulations such as GDPR can be explicitly recorded.

Challenges for Edge-AI

Undoubtedly, the greatest challenge for *Edge-AI* is in obtaining the large datasets that are needed to train deep neural networks. Data
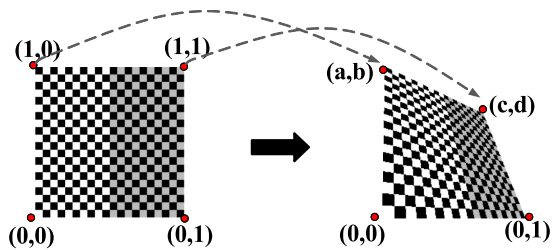
Figure 6. Novel data augmentation strategies can help grow the available training dataset for a specific problem; in this example, we show how iris data samples can be transformed to solve off-axis iris segmentation.[15]

acquisition is time consuming, and every problem needs a *ground truth* to train against!

Taking the iris segmentation problem as an example, every iris image should be marked-up with a ground truth. Even where some automated mark-up is possible, there must be a manual check to detect mark-up failures. This is very time-consuming, and the costs can quickly mount up for large datasets. This data bottleneck is a problem for all deep learning researchers, but more so for CE engineers who may need to adapt networks for multiple device models or in new use-case geometries such as off-axis iris authentication.

There are approaches that can help here, such as data augmentation (Figure 6) where a seed-dataset is modified and transformed in particular ways to match the underlying problem, as is done in the case of iris segmentation.[14,15] It has also been shown that we can train a network to learn how to make "new" data by combining existing samples in its convolutional layers.[55] Another technique that pairs two deep-learning networks in a configuration known as a generative adversarial network enables researchers to train a *data generator* that learns the key features of an existing dataset and can thus make "new" random data samples that match these.[56]

Data curation is another major challenge. As mentioned in "Device-Specific Aspects," if data samples are not carefully chosen to match the problem at hand, then neural networks can easily learn incorrect features from the training dataset. For AI applications in CE devices, this dependence of the solution on the training data is both a challenge and an opportunity. The better aligned the training

data is with the original sensor system, the more robust and accurate the trained network will be. For large research datasets, the training data are typically gathered by many devices and images are harvested from many online sources. If we consider the variety of video cameras used to create a collection of *youtube* videos, for example, it is easy to see that networks trained on such datasets will not be able to take account of device-specific characteristics.

Thus, the biggest challenge for *Edge-AI* is that of data. Improved tools and methodologies are needed to better support acquisition, annotation, and curation of training datasets. This is a fascinating topic and a follow-up article is planned to address it in more detail.

CONCLUDING THOUGHTS

A lot has happened in the last two years. Our research group has continued to work on a number of fascinating problems, leveraging deep learning techniques to achieve state-of-the-art solutions. In parallel, many other researchers have been working on and solving a broad range of problems in computer vision, machine learning, signal processing, data analytics, and advanced sensor fusion, all of which have relevance to new and emerging CE devices and services.

There has been explosive growth in the use of deep learning and advanced neural network methodologies, and much of this research can be leveraged into new CE solutions. The challenge today for CE engineers and researchers is how to pick and choose across this vast array of possibilities and deliver practical and useful solutions that can meet the needs of consumers.

ACKNOWLEDGMENTS

■ REFERENCES

1. J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision.," *IEEE Consum. Electron. Mag.*, vol. 6, no. 2, pp. 48–56, Apr. 2017.

2. P. Corcoran and S. K. Datta, "Mobile-edge computing and the internet of things for consumers: Extending cloud computing and services to the edge of the network," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 73–74, Oct. 2016.

3. P. Corcoran, "Mobile-edge computing and internet of things for consumers: Part II: Energy efficiency, connectivity, and economic development," *IEEE Consum. Electron. Mag.*, vol. 6, no. 1, pp. 51–52, Jan. 2017.

4. S. I. Venieris, A. Kouris, and C.-S. Bouganis, "Toolflows for mapping convolutional neural networks on FPGAs: A survey and future directions," *ACM Comput. Surv.*, vol. 51, no. 3, 2018, Art. no. 56.

5. A.-A. Erofei, C.-F. Druta, and C. Daniel Caleanu, "Embedded solutions for deep neural networks implementation," in *Proc. IEEE 12th Int. Symp. Appl. Comput. Intell. Inform.*, 2018, pp. 425–430.

6. M. Kotlar, D. Bojic, M. Punt, and V. Milutinovic, "A survey of deep neural networks: Deployment location and underlying hardware," in *Proc. 14th Symp. Neural Netw. Appl.*, 2018, pp. 1–6.

7. A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

8. Q. Lohmeyer, A. Schneider, C. Jordi, and J. Lange, "Expert opinion on drug delivery toward a new age of patient centricity? The application of eye-tracking to the development of connected self-injection systems," *Expert Opinion Drug Deliv.*, vol. 16, no. 2, pp. 163–175, 2019.

9. L. Cornish, A. Hill, M. S. Horswill, S. I. Becker, and M. O. Watson, "Eye-tracking reveals how observation chart design features affect the detection of patient deterioration: An experimental study," *Appl. Ergon.*, vol. 75, pp. 230–242, 2019.

10. M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth, and A. C. Trapp, "Detecting task demand via an eye tracking machine learning system," *Decis. Support Syst.*, vol. 116, pp. 91–101, 2019.

11. J. Choi, T.-H. Oh, and I. S. Kweon, "Human attention estimation for natural images: An automatic gaze refinement approach," *arXiv 1601.02852*, 2016.

12. J. Przybyło, E. Kańtoch, and P. Augustyniak, "Eyetracking-based assessment of affect-related decay of human performance in visual tasks," *Future Gener. Comput. Syst.*, vol. 92, pp. 504–515, 2019.

13. J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Trans. Consum. Electron.*, vol. 65, no. 2, pp. 179–187, May 2019.

14. S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end deep neural network for iris segmentation in unconstrained scenarios," *Neural Netw.*, vol. 106, pp. 79–95, Oct. 2018.

15. V. Varkarakis, S. Bazrafkan, and P. Corcoran, "Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets," *arXiv: 1903.00389*, 2019.

16. J. Lemley, A. Kar, and P. Corcoran, "Eye tracking in augmented spaces: A deep learning approach," in *Proc. IEEE Games, Entertainment, Media Conf.*, 2018, pp. 396–401.

17. P. Corcoran, "Biometrics and consumer electronics: A brave new world or the road to dystopia?" *IEEE Consum. Electron. Mag.*, vol. 2, no. 2, pp. 22–33, Apr. 2013.

18. P. Corcoran, "The battle for privacy in your pocket [notes from the editor]," *IEEE Consum. Electron. Mag.*, vol. 5, no. 3, pp. 3–36, Jul. 2016.

19. P. M. Corcoran, "A privacy framework for the internet of things," in *Proc. IEEE 3rd World Forum Internet Things*, 2017, pp. 13–18.

20. P. Corcoran and C. Costache, "Biometric technology and smartphones: A consideration of the practicalities of a broad adoption of biometrics and the likely impacts," *IEEE Consum. Electron. Mag.*, vol. 5, no. 2, pp. 70–78, Apr. 2016.

21. S. Thavalengal, P. Bigioi, and P. Corcoran, "Iris authentication in handheld devices—Considerations for constraint-free acquisition," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 245–253, May 2015.

22. J. J. Lozoya-Santos, R. A. Ramirez-Mendoza, S. Savaresi, J. C. Tudon-Martinez, and V. Sepúlveda-Arróniz, "Survey on biometry for cognitive automotive systems," *Cogn. Syst. Res.*, vol. 55, pp. 175–191, 2019.

23. A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Deeply vulnerable: A study of the robustness of face recognition to presentation attacks," *IET Biometrics*, vol. 7, no. 1, pp. 15–26, Jan. 2018.

24. S. Q. Liu, P. C. Yuen, X. Li, and G. Zhao, "Recent progress on face presentation attack detection of 3D mask attacks," in *Handbook of Biometric Anti-Spoofing. (Adv. in Computer Vision and Pattern Recognition)..* Cham, Switzerland: Springer, 2019, pp. 229–246.

25. S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, "Recent advances in face presentation attack detection," in *Handbook of Biometric Anti-Spoofing* (Adv. in Computer Vision and Pattern Recognition). Cham, Switzerland: Springer, 2019, pp. 207–228.

26. S. Thavalengal, T. Nedelcu, P. Bigioi, and P. Corcoran, "Iris liveness detection for next generation smartphones," *IEEE Trans. Consum. Electron.*, vol. 62, no. 2, pp. 95–102, May 2016.

27. A. F. Sequeira, S. Thavalengal, J. Ferryman, P. Corcoran, and J. S. Cardoso, "A realistic evaluation of iris presentation attack detection," in *Proc. 39th Int. Conf. Telecommun. Signal Process.,* 2016, pp. 660–664.

28. A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1031–1039.

29. Y. Zhao, F. Tang, W. Dong, F. Huang, and X. Zhang, "Joint face alignment and segmentation via deep multi-task learning," *Multimedia Tools Appl.*, vol. 78, pp. 13131–13148, 2018.

30. Y. A. U. Rehman, L. M. Po, and M. Liu, "LiveNet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Syst. Appl.*, vol. 108, pp. 159–169, 2018.

31. A. Sengur, Z. Akhtar, Y. Akbulut, S. Ekici, and U. Budak, "Deep feature extraction for face liveness detection," in *Proc. Int. Conf. Artif. Intell. Data Process.*, 2018, pp. 1–4.

32. Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou, W. Ou, and Y. Zhao, "Face liveness detection using convolutional-features fusion of real and deep network generated face images," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 574–582, 2019.

33. L. Wu, Y. Xu, M. Jian, X. Xu, and W. Qi, "Face liveness detection scheme with static and dynamic features," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 2, Mar. 2018, Art. no. 1840001.

34. V. Varkarakis, S. Bazrafkan, and P. Corcoran, "A deep learning approach to segmentation of distorted iris regions in head-mounted displays," in *Proc. IEEE Games, Entertainment, Media Conf.*, 2018, pp. 402–406.

35. M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 1151–1160.

36. M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3828–3836.

37. F. Husain, H. Schulz, B. Dellen, C. Torras, and S. Behnke, "Combining semantic and geometric features for object class segmentation of indoor scenes," *IEEE Robot. Autom. Lett.*, vol. 2, no. 1, pp. 49–55, Jan. 2017.

38. S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3479–3487.

39. S. Bazrafkan, H. Javidnia, and J. Lemley, "Semiparallel deep neural network hybrid architecture: First application on depth from monocular camera," *J. Electron. Imag.*, vol. 27, no. 4, 2018, Art. no. 043041.

40. C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Computer Vis.*, 2017, pp. 213–228.

41. B. E. Bayer, "Color imaging array," U.S. Patent 3 971 065, 1976.

42. P. Corcoran and P. Bigioi, "Consumer imaging I— Processing pipeline, focus and exposure," in *Handbook of Visual Display Technology*, J. Chen, W. Cranton, and M. Fihn, Eds. Berlin, Germany: Springer, 2014, pp. 1–25.

43. P. Amba, D. Alleysson, and M. Mermillod, "Demosaicing using dual layer feedforward neural network," in *Proc. Color Imag. Conf.*, vol. 2018, no. 1, 2019, pp. 211–218.

44. F. Kokkinos and S. Lefkimmiatis, "Deep image demosaicking using a cascade of convolutional residual denoising networks," *Lecture Notes Comput. Sci.*, vol. 11218, pp. 317–333, 2018.

45. M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 191.

46. W. Dong, M. Yuan, X. Li, and G. Shi, "Joint demosaicing and denoising with perceptual optimization on a generative adversarial network," arXiv*:* 1802.04723, 2018.

47. A. Buades and J. Duran, "Joint denoising and demosaicking of raw video sequences," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 2172–2176.

48. E. Schwartz, R. Giryes, and A. M. Bronstein, "Deep ISP: Toward learning an end-to-end image processing pipeline," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 912–923, Jan. 2019.

49. A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2016, pp. 1–6.

50. L. Bondi, L. Baroffio, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 259–263, Mar. 2017.

51. C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

52. *IEEE Games, Entertainment, Media Conf.*, 2018. [Online]. Available: https://twitter.com/i/moments/1061174127342559232

53. Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019.

54. S. Grzonkowski and P. Corcoran, "Sharing cloud services: User authentication for social enhancement of home networking," *IEEE Trans. Consum. Electron.*, vol. 57, no. 3, pp. 1424–1432, Aug. 2011.

55. J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

56. S. Bazrafkan, H. Javidnia, and P. Corcoran, "Latent space mapping for generation of object elements with corresponding data annotation," *Pattern Recognit. Lett.*, vol. 116, no. 1, pp. 179–186, 2018.

Peter Corcoran a Professor with the National University of Ireland, Galway, Ireland. He is an IEEE Fellow. Contact him at peter.corcoran@nuigalway.ie.

Joseph Lemley is currently working toward the Ph.D. degree with the National University of Ireland, Galway, Ireland. Contact him at j.lemley2@nuigalway.ie.

Claudia Costache is a Postdoctoral Researcher with the National University of Ireland, Galway, Ireland. Contact her at claudia.iancucostache@nuigalway.ie.

Viktor Varkarakis is currently working toward the Ph.D. degree with the National University of Ireland, Galway, Ireland. Contact him at v.varkarakis1 @nuigalway.ie.

Appendix J

Deep Learning for Consumer Devices and Services 3 – Getting More From Your Datasets With Data Augmentation

Deep Learning for Consumer Devices and Services 3—Getting More From Your Datasets With Data Augmentation

Peter Corcoran
National University of Ireland Galway

Claudia Costacke
National University of Ireland Galway

Viktor Varkarakis
National University of Ireland Galway

Joseph Lemley
National University of Ireland Galway & Xperi Corporation, Galway

Abstract—A key aspect to developing and successfully deploying neural network (NN)-based solutions is the availability of suitable datasets. In this article some of the challenges to acquire and annotate data are discussed in the context of new consumer devices. To increase the sample size of training data several approaches to augment a seed dataset are explained and discussed including a number of advanced, problem-specific techniques. A basic introduction to the concept of learned data augmentation is also provided.

■ **THIS ARTICLE FOLLOWS-ON** from an earlier publication[1] on the topic of Deep Learning and how it can be successfully applied to solve problems in today's consumer electronics devices. In this article, we will focus on the importance of training data and explain the importance of building a relevant dataset for a particular task or problem.

While deep learning methodologies have been proving themselves in many different contexts, it

is easy to get carried away and ignore some practical realities about deep neural network (DNN)-based solutions. First, a network is only as good as the data used to train it, and training an effective neural network (NN) requires a large annotated dataset. The challenges of building a dataset will be discussed later. Second, as engineers, we do not have detailed control over what an NN learns from a dataset—NNs can easily learn "bad" features that can lead to challenging problems when they are deployed in the field. Third, NNs can be sensitive to features that are not apparent to human observers; as an example, an image dataset acquired from a particular camera module may learn features specific to that imaging system, and may not perform adequately if the optics, the sensor, or even the image processing pipeline are changed.

Thus, while the NN-based approaches can perform with very high levels of accuracy in many contexts, they need to be approached with respect and understanding of how the NN works, and how it can be tuned and adapted for deployments in CE devices and services.

One key aspect to developing and successfully deploying the NN-based solutions is the availability of suitable datasets and, in this article, we will discuss some of the challenges of acquiring and annotating data, and more importantly, we focus on approaches to build larger training datasets by augmentation of the base dataset.

CHALLENGES OF BUILDING A DATASET

Data are a valuable commodity in industry today. Indeed, many companies rely on data for their core business and there has been much recent press attention on the thorny topic of personal data.[8] In the context of the data that we are interested in for today's consumer electronics applications, it is more useful to think in terms of image and video data and most of the examples discussed in this article rely on such data.

Today's DNNs require large datasets in order to converge on an accurate representation of a multidimensional data distribution. The deeper and more sophisticated the network, the greater the need for data. But, acquiring good quality data is a complex and costly process, and accurately annotating the data to match a particular problem adds further cost and complexity to the process.

Typically, to achieve sufficiently robust results that can be implemented in a consumer electronics imaging system, a dataset of at least 10 000 images is needed and many research datasets now exceed 1 000 000 image samples. Acquiring such large numbers of image samples is a challenge in itself, but when the problem of image annotation to provide a ground truth for training is considered the costs can quickly explode. Even where some automation can be introduced into the annotation process, there is still a need to check the accuracy of a significant sample of the dataset.

As a very simplistic example, consider the problem of building a personalization system for a consumer device that can distinguish a small set of individuals suitable for a group of users in an extended family. The goal is simple—match a person with one of up to 20 people registered on the system. Assume that we are doing this from scratch, so you will need to gather data samples to train the system—at least one sample of each user, but that would not be sufficient. Some people may always smile when you take their picture, others may scowl, so the system could learn to classify people by their facial expression rather than by other characteristics. Children will have smaller faces than adults, so the system might learn to distinguish them by facial size rather than by facial features. If some samples are acquired indoors in low lighting and others are acquired in daylight, the system might learn these characteristics instead. But, you cannot ask people to provide dozens of data samples to register with such a system because that will ensure that no one will ever use it!

So, how can the challenge of dataset creation be solved, or at least simplified? One key element in providing a practical solution is to expand a smaller dataset through a process known as data-augmentation. If this is undertaken in a careful and considered manner, it becomes feasible to grow a relatively small seed dataset of a few thousand images into a much larger dataset that can achieve high levels of generalization.

DATA ACQUISITION CHALLENGES FOR CONSUMER DEVICES

The starting point for image acquisition in a consumer device is the camera module. Each

camera has its own individual characteristics, derived from a combination of its optics, the CMOS image sensor, and the digital processing of the sensor data that typically yields a 3-channel RGB output image. You can read elsewhere about the complexities of the image acquisition pipeline,[4] but for the present discussion, we can consider the camera module as a unified subsystem.

Now, it would be nice if we could assume that the data samples obtained from any well-calibrated imaging subsystem are consistent across a range of external acquisition conditions. In other words, the output images from all RGB cameras provide similar data samples across the same range of lighting, focus, and exposure settings. Unfortunately, as any imaging or optical expert will quickly tell you, there are so many tradeoffs involved, which is simply not the case. Every camera is quite unique and its data samples will vary in unique ways.

It is not uncommon for a particular model of smartphone to have different camera modules in different geographic regions or for a new and improved camera module—in the CE industry that typically means a "cheaper" module—to be introduced into the manufacturing process. And from our experience, when training with a limited dataset obtained from a specific consumer imaging module, the resulting networks can be tuned to high levels of performance, but are typically highly dependent on that module.

As a consequence, a key challenge when introducing deep learning technology into consumer vision applications is to build training datasets that can be regenerated when new acquisition systems are introduced into production. Data augmentation can play a key role in simplifying this regeneration process.

Let us start by considering how some simple data augmentations can improve a basic dataset.

WHAT IS DATA AUGMENTATION?

The quantity of high-quality annotated data for a problem is often limited by the cost and complexity of acquiring such data. And even when an abundance of annotated data is available, there are risks that the methods or conditions of data acquisition, or even the data acquisition system itself, may influence the training process. In a nutshell, it can be challenging to predict what features the training process may extract from a dataset. Thus, there are examples

of trained networks that classify certain random noise patterns as animals or objects,[2] and others where inverting a single pixel can change the detected object class.[3] These are called adversarial examples and have been a subject of heavy research. Incidentally, using these adversarial examples for data augmentation has been shown to increase network robustness.

The reality is that any training dataset provides a finite set of samples of a much larger data distribution. But, it is possible to expand the original set of samples in a variety of ways that improve the training dataset and enhance the ability of the trained network to generalize. This process is known as data augmentation and can be best understood through some practical examples.

Examples of Basic Data Augmentation

To provide a context for our discussion of basic data augmentation, consider one of the most common objects that computer vision applications seek to analyze—the human face. Facial analysis encompasses many facets from the basic detection of a facial region, to analyzing facial expressions, features, tracking eye and lip movements, and distinguishing a particular individual from others.

To progress any facial analysis, a training dataset is required, but it is clear that it will be very limited in scope. Faces have a broadly similar shape and structure, but they can appear in a wide variety of poses, illumination conditions, and with different expressions. Given a limited facial dataset, how might we expand this in ways that could improve the performance of a trained network?

One simple generalization is to resize the facial samples that are available. This will enable larger and smaller faces to be identified by the trained network. Another simple generatization is to rotate the available facial samples—if every face is only provided in an upright frontal pose, then the network will learn that faces only exist as mugshots. Depending on the intended application you might limit rotations to a relatively small range, but if you want to detect faces in all possible orientations, then a selection of samples from the seed dataset should be rotated through a full 360°.

Another common augmentation is to add Gaussian noise to the data samples. Deep networks can learn subtle features or pattens from a dataset that are imperceptible to the human

observer; adding random noise to a selection of image samples disrupts many of these hidden patterns, ensuring that the network will be robust to learning such features.

It is also worth noting that augmentations can be combined. As an example, it is helpful to blur some of the image samples, so that a network can generalize to images that are not in sharp focus; it is also good to flip images to remove a left/right bias—these and other augmentations can be combined to further expand the original seed dataset. Figure 1 illustrates all of the above basic data augmentations applied to a single facial data sample. A network trained with such an expanded dataset will be better able to generalize to faces of different sizes, in different orientations and at various levels of defocus.

EXAMPLES OF ADVANCED DATA AUGMENTATIONS

When deep learning techniques are applied to solve practical CE problems, there may aspects of the specific problem-at-hand that can be amenable to more advanced data augmentations. These will be quite problem specific and can be best understood through examples. Here, we provide two related case studies, starting with the problem of iris segmentation for mobile devices. This leads, in turn, to the related problem of off-axis iris segmentation as found on the next-generation AR headsets.

Iris Segmentation of Low-Quality Smartphone Images

The subsequent feature extraction and pattern matching stages of an authentication workflow rely on the accurate segmentation of the iris. The failed segmentations represent the single largest source of error in the iris authentication workflow.[5] For an accurate segmentation, the exact iris boundaries at pupil and sclera have to be obtained, the occluding eyelids have to be detected, and reflections have to be removed, or flagged. Iris segmentation can be formulated as a binary classification problem—each image pixel is part of the iris region, or not.

Acquiring iris data for a mobile device is challenging. Iris sizes are typically in the range of 80–100 pixels, and as many images are blurred or of poor quality, and it is challenging to accurately annotate the iris regions. While some test
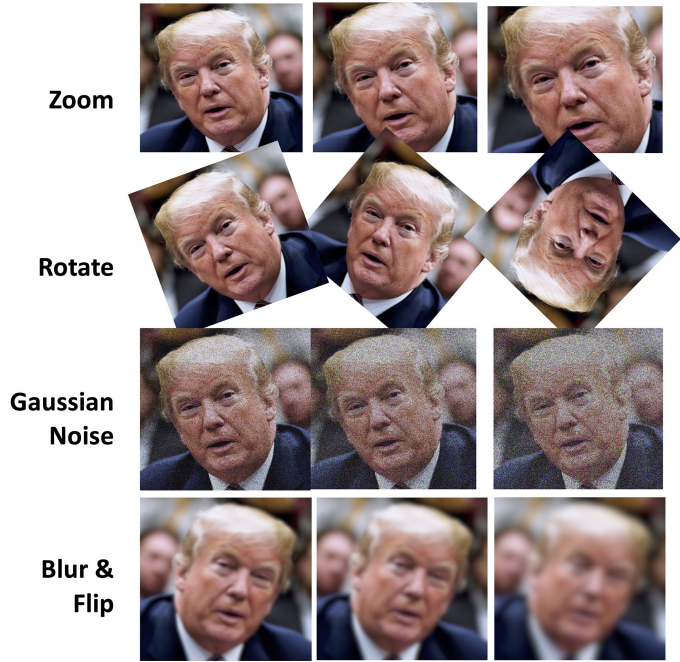

Figure 1. Some basic data augmentations in common use.

datasets exist, they do not provide sufficient training samples, and so researchers have explored the use of larger high-quality iris datasets (iris diameter > 300 pixels) to provide seed data.[5] High quality segmentation algorithms can provide an accurate ground truth without requiring manual annotation, and images can be augmented to mimic the low-quality images obtained from a handheld smartphone.

A range of augmentations are employed. All the high-quality images are downsized to provide iris regions with width of 80–100 pixels, typical for a smartphone camera acquisition. Image contrast
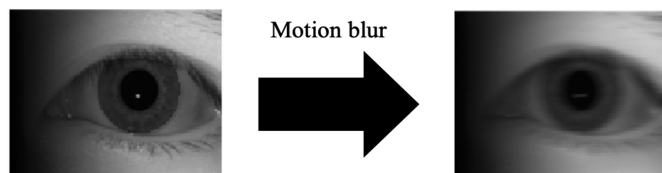
• **Shadowing**
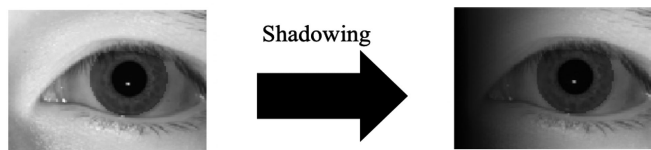
• **Motion Blur**

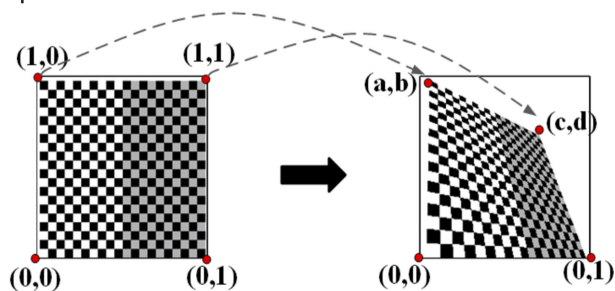**Figure 2.** Eye region firstly with shadowing applied, followed by motion blur.

Figure 3. Novel data augmentation strategies can help grow the available training dataset for a specific problem; in this example, we show how iris data samples can be transformed to solve off-axis iris segmentation.[7]

is also reduced by histogram remapping, but the most interesting augmentations seek to mimic the shadowing of the eye region that arises due to the wide variations in lighting conditions for smartphone acquisitions and the introduction of motion blur as the acquisition device is handheld.

These key augmentations are illustrated in Figure 2. A range of shadowing and motion blur are combined in[5] to provide a broad variation in the low-quality images introduced into the dataset. These advanced augmentations enabled the trained network to achieve state-of-the art performance on the available test datasets UBRIS and MobBio, both captured on actual smartphones.

Iris Segmentation for Augmented Reality Headsets

Related to the problem of iris segmentation on handheld devices, such as smartphones, is the
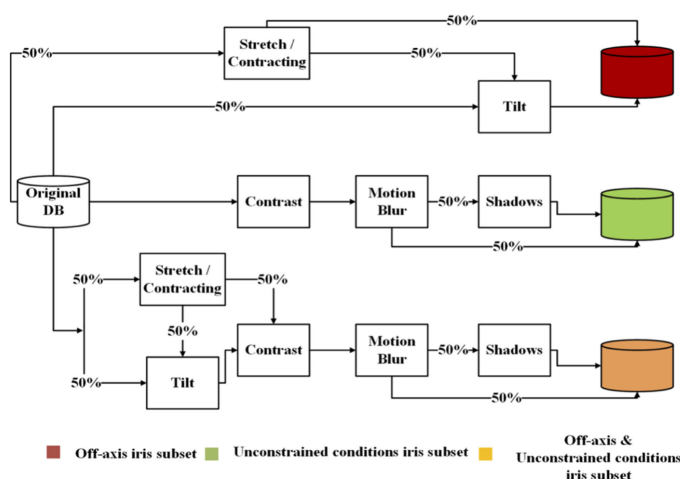

Figure 4. Augmentation workflow to generate training datasets for (a) off-axis iris regions, (b) unconstrained (smartphone) iris regions, and (c) a mixed dataset combining off-axis and unconstrained datasets.

emerging use of user-facing cameras on the next generation of augmented reality (AR) headsets, whereas the use of such cameras may initially appear counter-intuitive, they are required to track the eye-gaze of the user so that AR constructs can be rendered correctly onto the user's field of view. And eye-gaze tracking, as with iris authentication, requires the accurate segmentation of the iris and pupil regions within the eye.[6]

In this problem, we find that an affine transformation is required in addition to the augmentations used for iris segmentation on smartphones. This is illustrated in Figure 3.

This second example use case illustrates the potential of data augmentation in consumer device use cases. Starting from existing high-quality datasets that were originally intended for testing dedicated iris biometric authentication systems, it is now possible to generate a range of training datasets that can be applied across a range of new applications.

The high-quality seed dataset can thus be applied to train new iris segmentation, pupil segmentation, and eye-gaze analysis networks. Figure 4 illustrates three alternative data augmentation workflows that can be used to create three differing training datasets. Further tuning and customization of training datasets can be achieved by adjusting the shadowing and blur augmentations to closely match the characteristics of a particular acquisition subsystem. And if the user-facing camera positions are restricted, then the stretching and tilting ranges can be adjusted accordingly.

Some practical examples of off-axis iris segmentations are illustrated in Figure 5.

WHAT IS NEXT FOR DATA AUGMENTATION?

When we last considered the state-of-the art for machine learning, artificial intelligence, and computer vision in the context of consumer electronics systems, it was clear that the new hardware projects working to deliver new low-power neural network accelerators.

Learnable-Augmentation

The concept of learnable data augmentation is quite a recent development and appears to have originated in early 2017. This article describes a fully learnable data augmentation, where all components of the augmentation pipeline are learned using an artificial neural
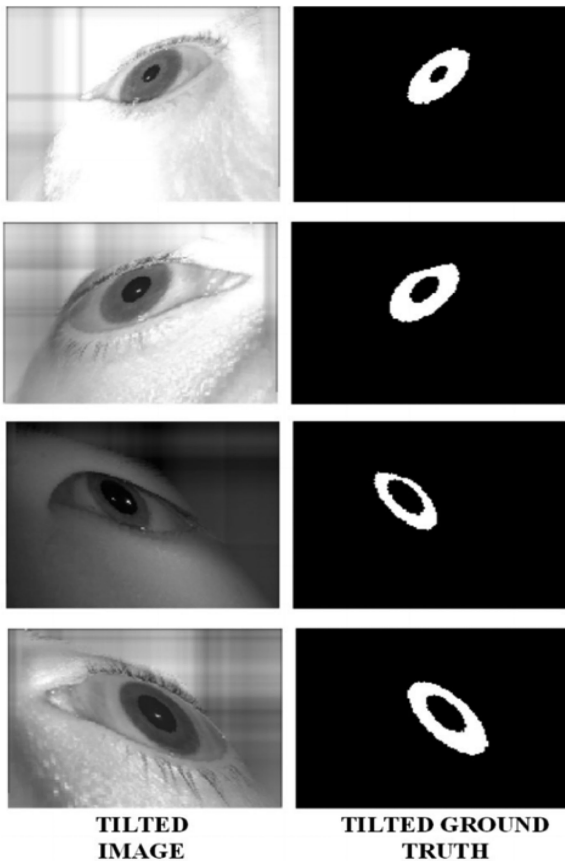
Figure 5. Off-axis iris regions; accurate, per-pixel, segmentation is essential for practical user authentication.

Figure 6. Images in the red box are created by a learned combination of the previous two images in that row.

network, *learnable data augmentation* as "Smart Augmentation."[7] This article addressed the questions: "Can a neural network learn to perform the labour-intensive process of data augmentation? Can a network or a collection of networks learn not just how to perform a task, but also how to create new data that will help it learn that task?"

The authors demonstrated an improvement of 5%–7% in the accuracy across a wide range of classification tasks. They also showed some augmentations that originate in the convolutional layers of the network, but that combine two different data samples in a way that makes sense to a human observer as shown in Figure 6.

Since 2017, there have been many works on learnable data augmentation, and these are detailed in a companion article, "Learnable Data Augmentation—Advanced Strategies for Improved Training of Deep Neural Networks." This companion article discusses these latest techniques in the context of CE devices and systems, and will be useful for readers who wish to explore more advanced data augmentation approaches.

CONCLUDING THOUGHTS

The goal of this article is to introduce the reader to the concept of *data augmentation* as it is applied in the field of machine learning, today. This powerful tool is widely used today to enable engineers and researchers to build working solutions with small problem-specific datasets.

This can be particularly helpful for engineering working in the CE field where it is often desirable to build a solution for a specific device or system, and where data gathered from other contexts may not be very useful, especially for fine-tuning the solution to achieve optimal performance. It is hoped for that the overview and examples given here will encourage other researchers to experiment more widely with the use of data augmentation to build larger and more effective training datasets.

■ REFERENCES

1. J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision," *IEEE Consum. Electron. Mag.*, vol. 6, no. 2, pp. 48–56, Apr. 2017.

2. M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols," *Pattern Recognit. Lett.*, vol. 57, pp. 17–23, 2015.

3. J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," (2017). Accessed: Jun. 28, 2018. [Online]. Available: http://arxiv.org/abs/1710.08864

4. P. Corcoran and P. Bigioi, "Consumer imaging I—Processing pipeline, focus and exposure," in *Handbook of Visual Display Technology*, J. Chen, W. Cranton, and M. Fihn, Eds. Berlin, Germany: Springer, 2014, pp. 1–25.

5. S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end deep neural network for iris segmentation in unconstrained scenarios," *Neural Netw.*, vol. 106, pp. 79–95, Oct. 2018.

6. N. Panigrahi, K. Lavu, S. K. Gorijala, P. Corcoran, and S. P. Mohanty, "A method for localizing the eye pupil for point-of-gaze estimation," *IEEE Potentials*, vol. 38, no. 1, pp. 37–42, Jan. 2019.

7. J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

8. The Cambridge Analytica Files, https://www.theguardian.com/news/series/cambridge-analytica-files, Last Accessed on 24 Mar 2020.

Peter Corcoran is an IEEE Fellow and professor at the National University of Ireland, Galway. Contact him at: peter.corcoran@nuigalway.ie.

Claudia Costache is a postdoctoral researcher at NUI Galway. Contact her at: claudancucostache@nuigalway.ie.

Viktor Varkarakis is a Ph.D. researcher at NUI Galway. Contact him at: v.varkarakis1@nuigalway.ie.

Joseph Lemley is a Ph.D. researcher at NUI Galway and a senior research engineer at Xperi corporation, Galway. Contact him at: joe.lemley@xperi.com.