DIFFUSE Project Report

Martin Torstensson, Leon Sütfeld,

Lorenzo Mirante, Jonathan Bergqvist, Svitlana Finér

Elektronik, mjukvara och kommunikation - FFI

2024-10-03

Vinnova Diarienr: 2021-05038

DIFFUSE – Disentanglement of Features for Utilization in Systematic Evaluation

1 Sammanfattning

När det gäller tillämpningar för maskininlärning (ML) som syftar till att förbättra trafiksäkerheten, komforten och tryggheten kan betydelsen av omfattande dataset för träning och utvärdering av prestanda inte överskattas. För att uppnå robusthet i dessa ML-modeller krävs dock dataset som omfattar en mängd individer med olika etniska egenskaper, vilket säkerställer att olika personliga egenskaper kan representeras. Trots vikten av stora och varierande datamängder kvarstår dock de inneboende begränsningarna för datatillämpbarhet; till exempel kan data som samlas in i en region inte nödvändigtvis generaliseras till en annan, vilket innebär utmaningar i olika dimensioner av dataanvändning.

För att ta itu med dessa utmaningar krävs innovativa metoder. Ett sätt är att samla in nya data, vilket medför avsevärda kostnader och bristande kontroll över vilka data som samlas in. Ett annat alternativ är att generera data som är skräddarsydd för tränings- och utvärderingsändamål, vilket är en strategi värd att överväga. En betydande utmaning uppstår dock under bildgenereringen - det är inte alltid enkelt att identifiera de specifika faktorer i en bild som bidrar till en modell.

Som svar på dessa utmaningar erbjuder vår föreslagna lösning ett nytt tillvägagångssätt: att utnyttja maskininlärningsmodeller för att generera syntetiska ansikten med olika och kontrollerbara ansiktsattribut och att återskapa attributen från bilderna. Detta tillvägagångssätt ökar inte bara mångfalden och volymen av tillgängliga data utan underlättar också förbättrad förklarbarhet, kontroll under autentiseringsprocesser och möjlighet att verifiera andra dataset. Genom att syntetisera data på det här sättet övervinner vi delvis begränsningarna i traditionella datainsamlings- och utbildningsmetoder, vilket i slutändan ökar effektiviteten och tillförlitligheten i ML-applikationer.

2 Executive summary in English

In the realm of machine learning (ML) applications aimed at enhancing traffic safety, comfort, and security, the significance of extensive datasets for training and evaluating performance cannot be overstated. However, achieving robustness in these ML models necessitates datasets comprising a multitude of individuals with diverse ethnic characteristics, ensuring the incorporation of varied personal features. Yet, despite the importance of large and diverse datasets, the inherent limitations of data applicability persist; for instance, data collected in one region may not necessarily generalize to another, posing challenges across various dimensions of data utilization.

Addressing these challenges requires innovative methodologies. One approach involves the collection of new data, with the downside of considerable cost and lack of oversight of what data is captured. Alternatively, generating data tailored for training and evaluation purposes presents itself as a viable strategy. Nonetheless, a notable challenge arises during image generation – identifying the specific factors within an image that contribute to a model is not always straightforward.

In response to these challenges, our proposed solution is to leverage machine learning models to generate synthetic faces with diverse and controllable facial attributes. This approach not only enhances the diversity and volume of available data but also facilitates improved explainability, control during authentication processes and ability to verify other datasets. By synthesizing data in this manner, we partially overcome the limitations of traditional data collection and training methods, ultimately advancing the efficacy and reliability of ML applications.

3 Introduction

Machine learning algorithms have significant potential to address problems within the automotive sector, such as perception algorithms for autonomous driving or face authentication algorithms for driver verification. However, one major requirement for these applications is access to large, well-distributed datasets that encompass a variety of scenarios for effective model training. Collecting such datasets is challenging due to the volume of data required and the necessity for accurate annotations to provide ground truth for the models. Moreover, verifying that a dataset comprehensively covers all relevant scenarios is difficult, as finite datasets may not capture every necessary detail.

For instance, face authentication algorithms often rely on deep learning techniques, particularly convolutional neural networks (CNNs), to identify individuals based on facial images. This process involves encoding facial attributes into numerical vectors, which can be compared to determine identity. However, variations in lighting, camera angle, and face position can lead to different encodings for the same individual, necessitating robust error minimization techniques.

To effectively train face authentication algorithms, image datasets must include diverse conditions, such as varying angles, lighting and environments, as well as a broad representation of individuals across different ethnicities and genders. This diversity is essential to avoid biases and ensure the model performs accurately across various groups. Collecting such comprehensive datasets often requires extensive geographical coverage, significantly increasing the cost and effort involved. To address these challenges, the DIFFUSE project aims to explore synthetic data generation and editing as alternatives to traditional data collection methods. By creating controlled synthetic data, the project seeks to enhance and offer new insights into the data distributions of existing real datasets.

Integrating synthetic data into real datasets entails analysing existing data for imbalances and identifying underrepresented scenarios, allowing for targeted supplementation. This controlled environment enables precise data reproduction and detailed annotations, critical for supervised ML training. After creating the synthetic data, quality checks and comparisons with real data ensure fidelity, and thorough documentation is prepared for release, reflecting the iterative nature of synthetic dataset creation, which evolves based on feedback throughout the ML training process.

3.1 Previous work

Previous projects such as the DRAMA2 project have highlighted the necessity for efficiently collecting labelled datasets, prompting the exploration of simulated environments for data generation [1]. This project aims to enhance the control over specific facial features while maintaining realism and variation, building upon methodologies like StyleGANv3 [2] and BFM [3].

Project number: 2020-02915
Title: DRAMA2 - Driver and passenger Activity Mapping Simulator
Programme affiliation: FFI - *EMK*
Decision-making agency: Vinnova
Summary o*f results and conclusions:* The project DRAMA-2 has *developed an evaluation methodology that is used as a base for the evaluations in this DIFFUSE project.*

Table 1: Summary of the Drama2 project details.

4 Purpose and research questions

We explore more efficient approaches for automatically generating and enhancing large datasets of diverse facial images, each with distinct characteristics. These methods aim to generate training data with customizable distributions, allowing us to enrich a given dataset with diverse face shapes and features that the model needs to learn. At the same time, we ensure the data retains the essential properties of the real-world data space. The project focuses on investigating the use of disentangled and interpretable latent spaces within state-of-the-art generative frameworks, such as GANs and 3DMMs, to validate other machine learning models and generate synthetic data for training purposes. The primary emphasis will be on generating facial images, with driver face authentication being one of the application areas where these methods will be tested.

The project addresses the following main research question:

- **RQ1:** How can feature disentanglement aid in increasing control in the generation of datasets for evaluation purposes?

To support this pursuit three more research questions are proposed:

- **RQ2:** How can we improve upon the feature disentanglement of the current state-of-art methods?
- **RQ3:** How can we use feature disentanglement to train more explainable machine learning models?
- **RQ4:** How can we use feature disentanglement as a basis to encode and compare previously collected datasets?

5 Disentanglement of GAN-based models

5.1 Background

5.1.1 Advancements in Generative Modelling

In recent years, generative modelling has seen remarkable progress, with methodologies such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) [4][5]. GANs in particular have undergone significant advancements such as Progressive Growing GAN (PGGAN), StyleGAN, and BigGAN, leading to enhanced image quality, detail, and scalability [6][7][8][9][10]. In the GAN architecture, two neural networks act as adversaries a “minimax game”; the generator network creates images from random noise vectors, while the discriminator network tries to tell the generated images apart from real ones. In this process, the discriminator learns what distinguishes the images in the training dataset from the images generated by the generator, while the generator learns what properties its generated images need to have in order to trick the discriminator into recognizing them as images from the training set. In this way, both networks iteratively improve over the course of training – the generator gets better and better at producing realistic looking images while the discriminator gets better and better at detecting the remaining differences; ideally this process continues until the generated images become indistinguishable from real images [4].

Features in the images, e.g., facial features, facial expression, hairstyle, direction of light, and background objects in case of facial images, are associated with directions in the vector space of the random noise vectors at the input of the generator. This presents an opportunity for image editing: By changing the input vector slightly in any direction, features of the image should also change slightly. To make this approach useful and powerful in practice, one would like to find all relevant features isolated from one another in orthogonal directions in feature space; one could then manipulate them individually and without inadvertently changing other features in the image as well. This would represent a perfectly disentangled and complete latent space where each distinct image feature maps onto one direction in the latent space and vice versa. In practice however, this is not usually the case: Features are typically located on non-orthogonal directions in latent space, making it impossible change one feature without changing others as well. Moreover, the latent space is often curved: Linear interpolation between two images in latent space should ideally yield a series of images that smoothly interpolate between the features found in the two images. For instance, all linear interpolations between the latent representations of two photos of 30-year-old people should also yield roughly 30-year-old people with a mixture of the original facial features. In practice, such interpolations often yield photos of different age-groups and with facial features not present in any of the original photos [11]. These issues to date prevent the use of these techniques for targeted image editing or the creation of new images with controlled features.

Techniques like InfoGAN and modified StyleGAN approaches have addressed disentanglement through unsupervised approaches and mutual information maximization [12][13], with recent studies exploring regularization methods such as contrastive learning for effective disentanglement [14][15]. Examples of images generated with StyleGANv3 [2] can be found in Figure 1.

Figure 1: Example images from StyleGAN v3 (image courtesy [2]).

5.1.2 Entangled latent spaces

By default, GAN models often show moderately to highly entangled latent spaces with non-orthogonal directions of change for individual features and curved and irregularly warped spaces preventing the manipulation of isolated features and smooth interpolation between images. To address these issues, [12] propose a GAN model that learns meaningful and disentangled representations by maximizing the mutual information between a fixed small subset of the GAN's input noise vectors and the generated images. They decompose the input noise vector into a source of noise z and a latent code c that represents the structured semantic features of the data distribution. They propose a new information-regularized minimax game that includes the standard GAN loss and the mutual information between c and G(z, c). In order to minimize the loss, they use an auxiliary distribution Q(c|x) that is parameterized as a neural network, i.e., a classifier network trained to extract the state of input vector c from generated images x. This approach necessitates the network to make the state of the dimensions in c visually detectable and encourages the use of easily distinguishable and isolated features to express the state of c visually. While helpful for the disentanglement of at least a few feature dimensions, this approach leaves large parts of the input vector unattended, and thus potentially entangled. Additionally, the training is unsupervised, meaning that while no labels are needed it is also not enforcing any specific choice of attributes to be controlled with the c vector.
[16] analyze the latent space of StyleGAN2 specifically. They compare disentanglement (i.e., each latent dimension controlling a single visual attribute) and completeness (i.e., each visual attribute being controlled by a single dimension) of three spaces – Z, W and S. Z is the typically normally distributed input space. The random noise vectors $z \in Z$ are transformed into an intermediate latent space W via a series of fully connected layers. Each $w \in W$ is further transformed to channel-wise style parameters s, using a different learned affine transformation for each layer. The space spanned by these style parameters is referred to as S. In their experiments, the authors show that S scores highest in terms of disentanglement and completeness.
[17] proposed a closed-form factorization method that is able to discover the latent semantic directions learned by GANs. They base their idea on the fact that GANs project a latent code (a noise vector z) to a photo-realistic image step by step, where each step learns a projection from one space to another. The method explores the first projection step that directly acts on the latent space, however,

they claim that their method can be applied to any layer. Their experiments show that this approach can be easily applied to StyleGAN, BigGAN, and StyleGAN2.

5.1.3 Image editing with GAN inversion

In order to be able to not only edit images created from random noise vectors, but also edit existing images, researchers are exploring the mapping of images back into the latent space of GANs in a process named GAN inversion. GAN inversion maps a real image x to the latent space, thus producing a latent vector that ideally leads to a faithful reconstruction of the image by the pre-existing generator. By varying the latent vector in different directions, one can then edit the corresponding attributes of the real image [18 if the latent space is well-disentangled. When used in practice this technique can be used to create new data with better control of the included features or to balance already existing datasets.

5.1.4 Bijective learnt transformations of latent spaces

[11] present an architecture called Latently Invertible Autoencoder (LIA) that enables better disentanglement of a latent space y by introducing a bijective, i.e., reversible network that decouples the latent space from the random noise vector z. The idea is that the default GAN architecture forces any dataset's feature distribution to be mapped onto a latent space with a fixed distribution (i.e., that of the random noise vector z), promoting or even enforcing entanglement of features with varying distributions. By decoupling the latent space y from the fixed distribution of noise vector z, their approach enables y to take on a distribution that is shaped by the distribution of features in the training data. In a second step, the authors train an encoder F that maps a given image x into the generator G's latent space such that the generator creates a faithful reconstruction of the original image: $G(F(x)) \approx x$ (GAN inversion). In choosing an invertible network for the decoupling, and training an additional encoder to map from image space x into latent space y, the authors create an autoencoder architecture with naturally decoupled latent space at the centre, while allowing for lossless mapping between z and y.
The bijective, i.e., invertible, property of the used NICE [19] network between z and y is achieved as follows: The input vector z is randomly divided into two equally sized parts z_1 and z_2, and a neural network f_1 transforms z_2 and the output is added to z_1 to form z_1'. A second neural network f_2 then transforms z_1' and the output is added to z_2 to form z_2':

$z_1' = z_1 + f_1(z_2)$
$z_2' = z_2 + f_2(z_1')$

Finally, z_1' and z_2' are concatenated to form the output of one layer in the NICE network. The inverse mapping from z' back to z is easily calculated as:

$z_2 = z_2' - f_2(z_1')$
$z_1 = z_1' - f_1(z_2)$

This process allows an arbitrary number of layers to be chained, each with an arbitrary split of the input vector in equal halves. While the expressivity of a NICE network is somewhat restricted compared to a classical fully connected network (multi-layer perceptron; MLP), the ability to compute the exact inverse without any additional training can enable novel network architectures and training approaches that would otherwise be impeded by imprecise inverse calculations and additional training requirements.

However, in the LIA approach, the bijectivity is not strictly required for the disentanglement – while it creates a neat symmetry in the architecture and enables lossless translations between the y and the z

space, it is not obvious why one would need to map back into z from the latent space y, as y is the disentangled space in which any image editing, mixing or interpolation would occur.

A regular fully connected network can equally enable the decoupling of z and y that facilitates the disentanglement of y and has in fact been implemented in the StyleGANv3 architecture, albeit under a slightly different naming convention; the more disentangled latent space is here referred to as style space w^+ instead of y.

5.2 Method

5.2.1 Structuring and explicit disentanglement of latent spaces

While StyleGANv3's latent space w^+ is understood to be somewhat disentangled due to its decoupling from the random noise vector z, it is not explicitly driven to maximize disentanglement or completeness and does not lend itself to controlled image manipulation or mixing as is. The goal in this project is to find and showcase a process that creates a latent space that not only maps individual visual features onto orthogonal directions, but also aligns these directions with the dimensions of the latent space. This means that each dimension, or cell, in the latent vector should represent – and manipulate – exactly one visual feature in the generated images. Such a mapping would not only allow for a straightforward assessment of which feature is encoded where in the latent space (making the network highly explainable) but would also allow for the analysis of any applicable dataset in terms of the frequency and distribution of features within it.

To approach this goal, we implemented a network architecture that uses a StyleGANv3 backbone pretrained on facial images as the generator and a matching pretrained encoder from [20] to map any given image into StyleGANv3's w^+ latent space. This represents an autoencoder-like circular architecture, enabling rather faithful and realistic looking reconstructions of facial images:

Let G be the generator pre-trained on a distribution X, F a matching encoder, and $x \in X$ an image from the original distribution, then $w^+ = F(x)$ is the latent representation of image x in StyleGANv3's latent space. Further, $x = G(w^+)$ is an image generated from the latent vector w^+ and $x' = G(F(x))$ the reconstruction of x.

In order to create a latent space that fulfils the disentanglement and completeness properties, we propose to use an invertible network T that maps from the latent space w^+ to a second latent space $\hat{w}$, and train T to structure the latent space in ways that enhance the disentanglement with respect to visual features in the image. To guarantee the stability of the existing latent space w^+ and with that the quality and diversity of the generated images, we freeze the weights of both the generator G and encoder F, and train only the translator network T, as well as discriminator D where applicable. This choice also makes the approach more easily applicable to problems involving the deliberate shaping of latent spaces in other architectures, as well as reducing the computational demand of the training procedure. The basic network architecture is illustrated in Figure 2.

The translation of convoluted, entangled, or generally unstructured latent spaces into structured and disentangled latent spaces is tightly linked to explainability and explainable AI, contributing to RQ3: It provides direct insights into how and where information is represented at the very core of neural networks, making internal states interpretable and providing helpful insights into algorithms otherwise often regarded to be black boxes. Invertibility of translator networks attached directly to representation layers further enables the precise manipulation of internal network states, providing a method to study the relation between internal states and network outputs. It should be noted that while this approach can in principle be applied to most deep neural networks, finding suitable ways of training such translator networks can often be challenging.

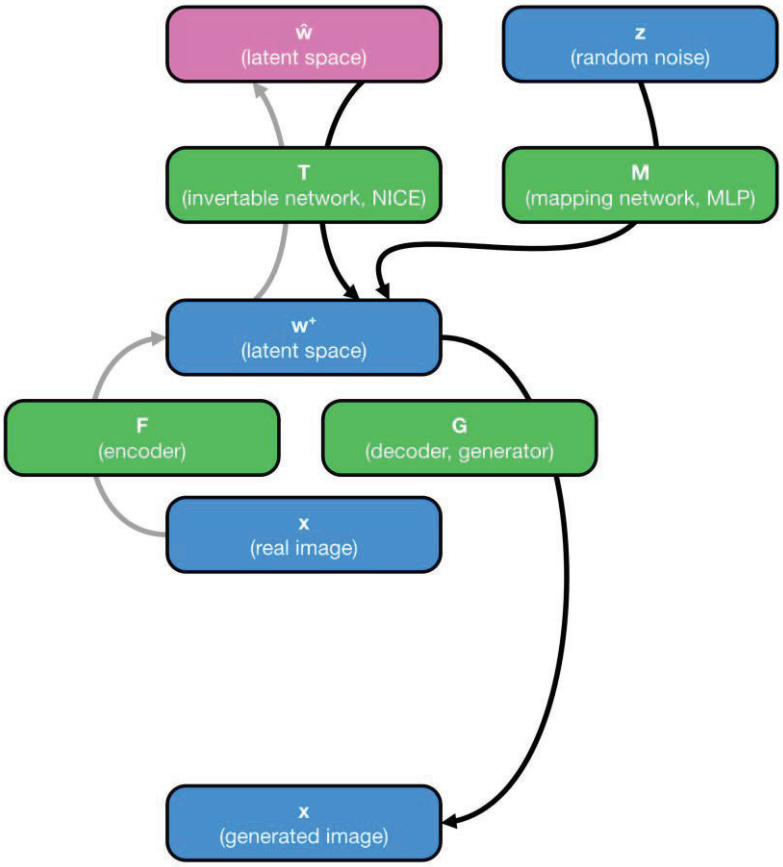

Figure 2: Illustration of the chosen network architecture. Functions (neural networks) indicated in green, data spaces in blue with the exception of the in-focus latent space ŵ (highlighted in pink). Gray arrows: Data flow for image embedding / analysis (inference). Black arrows: Data flow for image generation (inference).

For our use case of disentangling the latent spaces of GANs, training the translator network T is possible via a number of loss functions that are consistent with disentanglement properties. Each such loss function needs to relate image features with features of the latent space. Due to the circular design of the chosen architecture, we can employ both loss functions relating visual features of real images with their corresponding latent representations in ŵ, as well as loss functions relating properties of latent vectors ŵ with visual features of synthesized images. The ability to involve both the encoder and decoder in the training of T can be helpful in addressing stability-related issues that may occur during training.

Since the existence or extent of most visual features isn't straight forwardly observable from raw pixel values, we need to operationalize these features in some way. To this end, we have multiple options: The most straight forward way is to use a dataset with annotated visual features, relying on human appraisals of the images and accepting the provided selection of annotated attributes. Another approach is to use additional machine learning models pre-trained to detect relevant features in the images, e.g., features in facial images that enable identification taken from face recognition networks – an approach closer to the actual data, potentially providing a richer and more informative training signal for the translator network. However, training a network to directly associate specific dimensions of the latent space with specific features of the image space isn't the only way to achieve disentanglement and/or improve the suitability for image analysis, editing, and interpolation. We can

use additionally available data like identity labels or knowledge about the spatial extent and other properties of targeted features to structure the latent space in meaningful ways, even if no suitable annotations are available. For instance, for the augmentation of datasets of facial images, a separation of the latent space into one part containing identity-related features, and one part containing all features not pertaining to the identity of a person can be tremendously helpful: It would allow to show the same person in a variety of settings, or show a variety of people in a specific setting. Further subdivisions of the latent space, e.g., into features of greater or reduced spatial extent, or features specific to certain image regions can yield additional structure and reduce entanglement further.

At this point it is worth discussing the concept of visual features and disentanglement in a little more detail. In particular, we must recognize that perfect disentanglement of visual features is impossible for most data distributions. Subdividing the feature space into an endless string of miniscule and entirely independent features would not only require enormously large representation spaces to make room for all the individual features, but it would also defeat the purpose of providing a way to purposefully analyse and augment feature distributions in datasets. For these purposes, we require at least some features that agglomerate a number of sub-features within them. Gender, ethnicity, age, and facial expressions are examples of complex features that are composed of highly correlated sub-features, and yet make a lot of practical sense to be kept as single features, while also keeping many of the corresponding sub-features editable individually. For instance, skin tone may vary both as part of an ethnicity feature, but also independently of ethnicity, resulting in at least two features that are desirable to have but which can't even conceptually be orthogonal to one another. Despite the unattainability of perfect disentanglement, a lot of features in most target distributions are conceptually independent, giving ample reason to pursue their disentanglement in latent space.

In this project, we conducted two main lines of inquiry into disentanglement within the GAN-based approaches: One, by enforcing a precise matching of latent dimensions with select visual features using an annotated dataset, and two, by dividing the latent space into distinct regions for identity-related and identity-unrelated features, using only the ID-labels. We will now outline the specific implementations for these two approaches.

5.2.2 Dimension-specific disentanglement

As mentioned above, the circular layout of the chosen architecture and invertible translator network allows us to train the network in both directions; (1) while mapping from real images into $\hat{w}$, and (2) while mapping from $\hat{w}$ into synthesized images.

For the first of these, depicted in Figure 3, we utilize the encoder, the NICE network and the CelebA – HQ dataset, see Appendix A, training the translator network T to learn a mapping from the w^+ space into $\hat{w}$. The $\hat{w}$ space is divided into two subspaces: an attribute space that is intended to be a perfect recreation of the attributes in the dataset – this is the disentangled part of the space – and the remaining space that is not explicitly conditioned, available for information that is not related to the attributes. The approach uses a pair of an image and a vector containing binary annotations. The image is encoded, i.e., mapped into w^+ and further passed through T for a representation in $\hat{w}$. The attribute space representation is then trained to match, i.e., recreate and represent the annotations associated with the image via a MSE loss, referred to as encoder loss. This allows for the direct use of the attribute space as an analyzer for the existence of a feature in a given image. Applied to an entire dataset, the observed distribution of features in the attribute serves as an estimator of the distribution of features in the dataset.

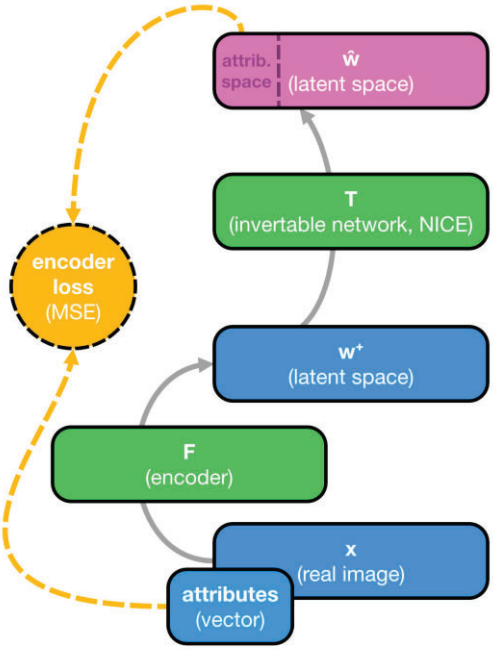

Figure 3: Illustration of the chosen encoder loss network architecture. Functions (neural networks) indicated in green, data spaces in blue with the exception of the in-focus latent space ŵ (highlighted in pink). Gray arrows: Data flow for image embedding / analysis (inference). Orange arrows: Data flow related to the training of T.

For the inverse direction, depicted in Figure 4, training starts with the generation of a random vector in ŵ and setting the attribute space to match the attributes of a random image from the CelebA – HQ dataset. This latent representation ŵ is then transformed into w^+ via T, which is in turn used by generator G to create an image. This image is processed by two additional networks:

(1) An attribute classifier network H pretrained to recreate the attributes from the visual information in the image, followed by an MSE loss between the recreated attributes H(x) and the original attributes from the attribute space in ŵ. The attribute classifier is based on a ResNet18 architecture trained on the larger CelebA dataset, see Appendix A. It takes images as inputs and extracts the associated attributes. Since the imbalances present in the distribution of many of the classes in the CelebA dataset can hinder training, two attributes with fairly balanced distributions were chosen: "Male" and "Smiling" should present good visual markers to evaluate the training method as a proof-of-concept. If it is possible for the network to successfully disentangle these two then the same process should be easily applicable for further attributes, provided proper care is given to the balancing of class distributions.

(2) Second, the pretrained StyleGANv3 discriminator, followed by a discriminator loss, used to make sure that the output of the generator produces realistic images. It is trained via adversarial training, classifying images into real and fake images. Two methods were tested; one with a frozen discriminator, and one in which the discriminator was trained alongside T.

The frozen discriminator and the frozen classifiers lead to poor output. In these cases, T is likely to overfit to the specific patterns the classifier and discriminator are looking for, resulting in unrealistic images. When training the discriminator alongside T, image quality is kept more stable as T cannot learn to replicate static patterns and is forced to learn better general representations.

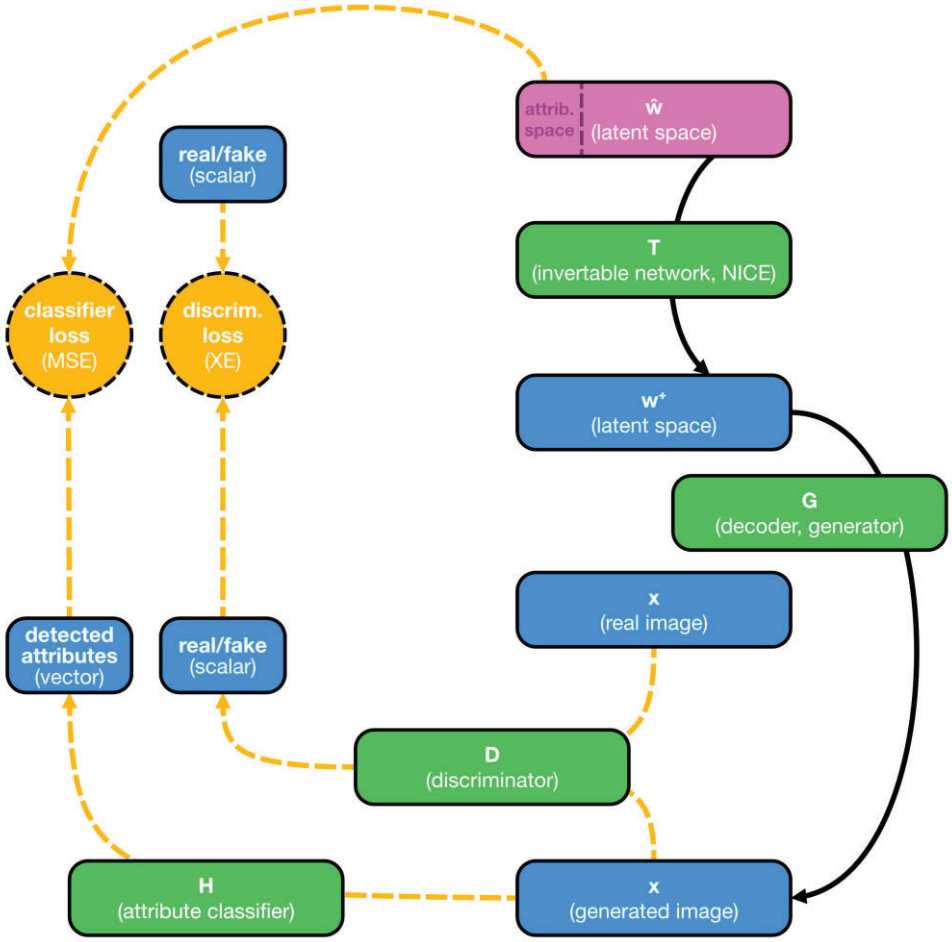

Figure 4: Illustration of the classifier and discriminator loss network architecture. Functions (neural networks) indicated in green, data spaces in blue with the exception of the in-focus latent space ŵ (highlighted in pink). Black arrows: Data flow for image generation (inference). Orange arrows: Data flow related to the training of T.

Finally, a third model merges both approaches outlined above, illustrated in Figure 5.This model has all the functionality of both individual approaches outlined above. It is the basis for the results shown in the result section and is contributing to all four RQs, the deliverables of disentanglement model in WP2 and the data generation model in WP4.

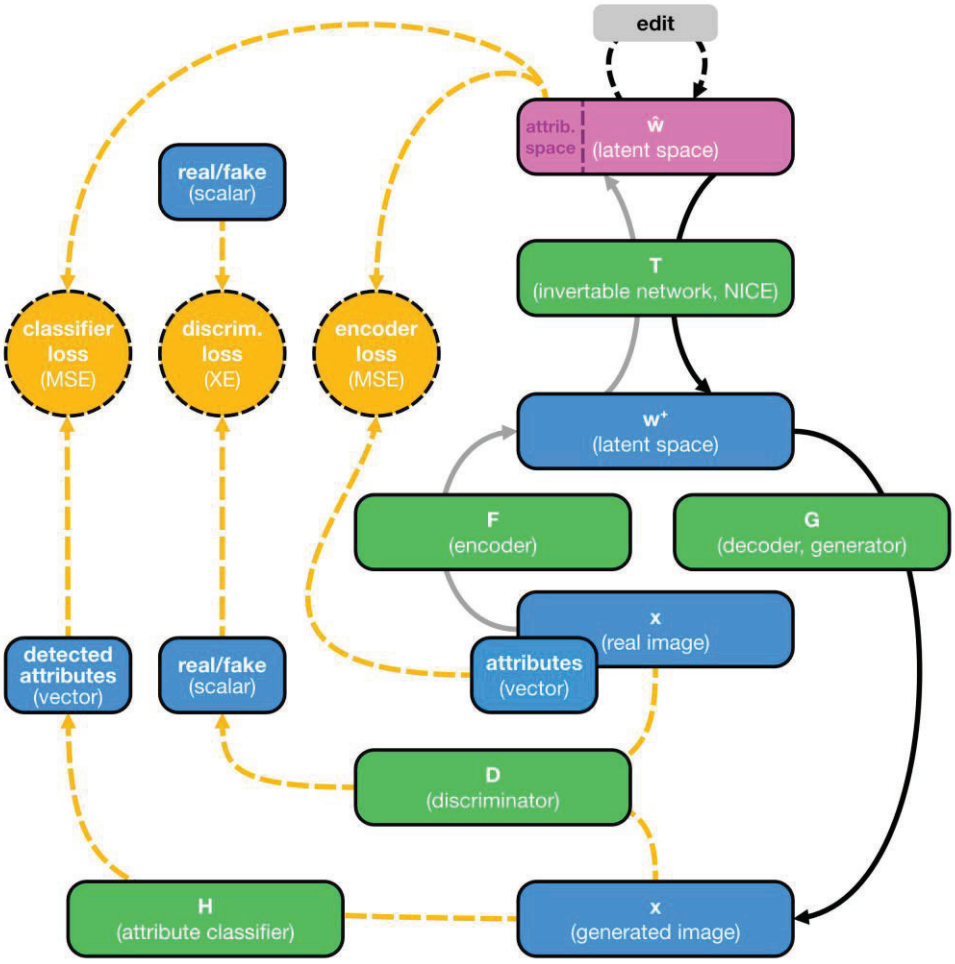

Figure 5: Illustration of the complete network architecture. Functions (neural networks) indicated in green, data spaces in blue with the exception of the in-focus latent space ŵ (highlighted in pink). Gray arrows: Data flow for image embedding / analysis. Black arrows: Data flow for image generation (inference). Orange arrows: Data flow related to the training of T.

5.2.3 Structured latent spaces and division of identity-related and unrelated sections

Our second line of inquiry into disentanglement of GAN-based approaches is to structure the latent space by division into distinct regions, specifically for identity-related and identity-unrelated features, using only the ID-labels available in the CelebA dataset. This approach can equally be seen as proof-of-concept for a more general structuring of the feature space via weak supervision signals. The same general approach could be used to arrange features by size (e.g. skin tones generally affecting larger regions of the image than the nose shape), image region (e.g., head hair features being located higher up than beard features), or number of separate image patches affected (e.g., eye-related features typically affecting two distinct regions while moth-related features only affect one). Once effective ways are found to perform such separations of the latent space, a combination of several of these approaches may yield high degrees of disentanglement with properly isolated features. The principal advantages of this approach over the previously introduced dimension-specific disentanglement are the missing dependence on annotated features, making this approach much more generally applicable to various datasets, as well as it affecting the entire latent space as opposed to a mere fraction of it.

However, we begin the inquiry in this project with a focus on the separation of identity-related and identity-unrelated regions in the latent space. This is possible, e.g., through the use of contrastive

learning: By comparing latent representations of two images at a time. In the identity-related section of the latent space, the representation of two such images should be very similar or identical if both images show the same person, and unrelated or distinctly different when showing different people. The identity-unrelated section of the latent space in this case should contain features like the background color or objects, lighting direction, the angle between face and camera, the facial expression, etc. We do not generally have a signal or label that we could use to establish similarity of these features, but we can be creative about how to amass these features in this part of the latent space. First, by amassing all identity-related features in the other section, the identity-unrelated features should naturally be forced out of the other and in into this section of the latent space. We can also create additional space for these features to occupy by actively forcing identity-related features out of this space. For instance, a face-identification network can be trained to identify faces based on the representation in this second section of the latent space, and an adversarial training regime can be employed to train the translator network T to make it harder for the face-identification network to successfully identify a person based on the representation in this identity-unrelated space. The only way to achieve this is by "hiding" all features giving hits towards a person's identity in the other section of the latent space. These are some examples of the ways we can train T to restructure the latent space in the desired way.

For the purposes of authentication methods, separating the latent space into an identity-related and an identity-unrelated part is a promising approach. Once trained to map different images of the same person to very close or identical places in the first section of the latent space, identification by distance/similarity (e.g., cosine similarity) in this space is expected to be highly accurate. This creates the basis for the authentication model in WP3 and provides a disentanglement-based method for how to approach and improve upon the solutions for this problem. While we discuss the principles and mechanisms behind this idea, unfortunately it was not possible to realize this authentication method within the frame of the project.

The approach of separating the latent space into an identity-related and unrelated part was within this project explored in a master thesis [21]. The following sections presents some of the main results achieved in this thesis. For more detailed information on the methodology and results, the reader is encouraged to read the thesis.

5.3 Results

In this chapter the results for image editing and feature reconstruction with the GAN-based method are presented contribution to the evaluations in WP5.

5.3.1 Separation of ID and non-ID features

This chapter contains an excerpt of results from the master thesis [21]. In Figure 6 and Figure 7 are two examples of mixing the $\hat{w}$ spaces of two individuals to illustrate what information goes into each part of the vector. In both cases the identity part of the $\hat{w}$ vector for the person in the first column is used and combined with the non-ID part for each person in the first row resulting in the remaining images in row 2. More information and analysis can be found in the thesis.

Figure 6. Mixing of image pairs: The first row shows the original images. The non-identity halves of the corresponding latent representations in W, trained with the latent distance objective, are mixed with the identity part of the image in the first column of the second row. All other images in the second row are the resulting mixed images. Excerpt from [21].*

Figure 7: Mixing of image pairs: The first row shows the original images. The non-identity halves of the corresponding latent representations in W, trained with the latent distance objective, are mixed with the identity part of the image in the first column of the second row. All other images in the second row are the resulting mixed images. Excerpt from [21].*

5.3.2 Attribute classifier

The classifier is intended to find the attributes in an image matching the annotations in the dataset. For the case of this study two attributes are used “male” and “smiling”. That means the network associates each image with two scalars one being a gradual scale for the binary classification for male and another for smile. During the training of the network the larger CelebA dataset was used as the success of the classifier finding good representations in the images to associate with the classes are paramount for the success of the rest of the network. If the dataset is too small there is a risk that the classifier overfit to specific features that are common in that dataset, which are not representative for the classes on a larger scale. Although this is an issue for all datasets additional data in many cases leads to a better generalization of the network. The resulting accuracy for the two classes can be seen in Figure 8 as a function of the number of epochs trained in the CelebA validation dataset. For the total score an accuracy of 96% is reached, where the male class has a higher accuracy than that of smile. This could be due to how there are more features in the images that are related to and changed with the male class compared to the smile class, which is quite restricted to a small area in the face. Small performance improvements can still be seen in the end of the training, and it is possible that longer training in combination with a lower learning rate could further improve the results.

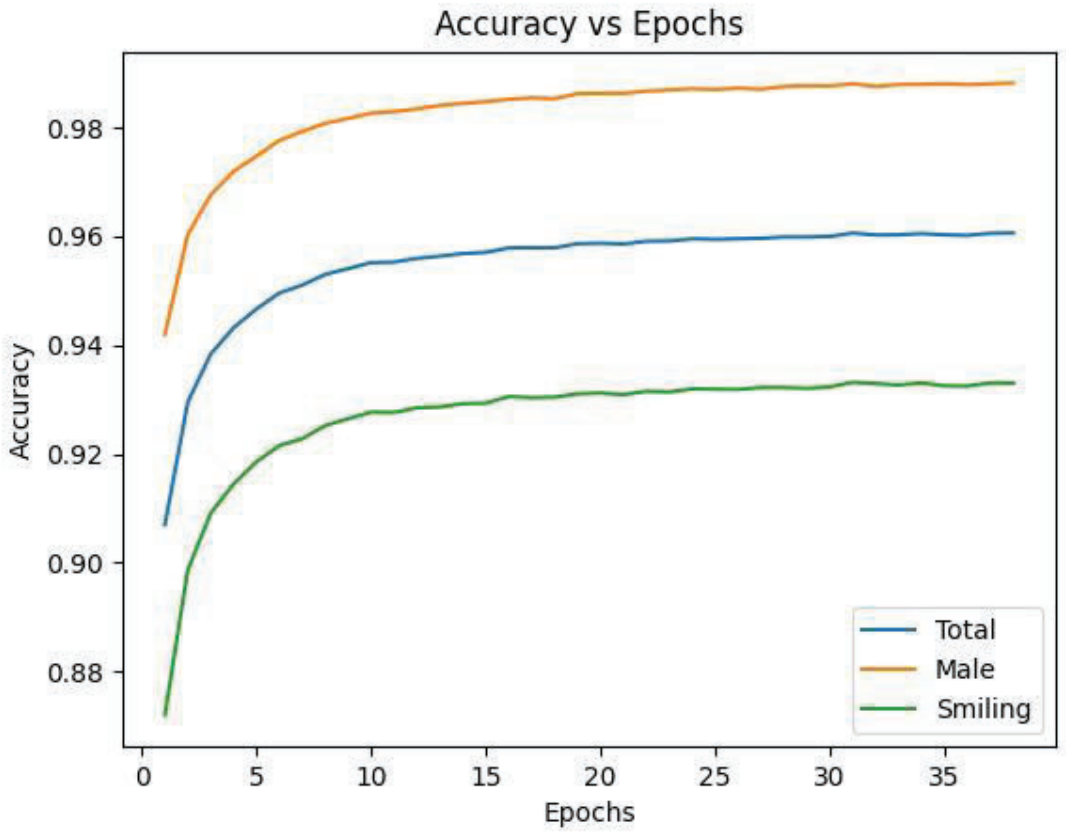

Figure 8: Accuracy change on the CelebA validation dataset during training for the classifier network.

5.3.3 Image editing with the GAN network

With the complete GAN architecture depicted in Figure 5 a series of experiments were done for the task of image editing. The goal of the task is to take a real image, modify it in a disentangled latent space and then generate a new image as similar as possible to the original while changing specific features. This was achieved by following a structure where an image is encoded into the w^+ space, converted into the $\hat{w}$ space through the NICE network, had one or more parameters modified, reverted back through the nice network into the w^+ space, and finally converted into an image with the generator. Two examples of how these modifications can look are shown in Figure 9 and Figure 10 one with an original image of a female and one of a male to show the networks capability to work in both directions. These figures show both the original image as well as a reconstructed image, which is how the image would have looked after encoding and generation without any modifications. Furthermore, each row depicts a sliding modification of what value is set in the $\hat{w}$ space for the specific attribute or attributes starting at –4 and ending at 4. Thus, the model attempts to create more feminine or lacking smiles in the left side and moving over to more masculine and more smiling images depending on the row.

Figure 9: First example output of the GAN based disentanglement network varying the aspects for male/female and smiling/not smiling.

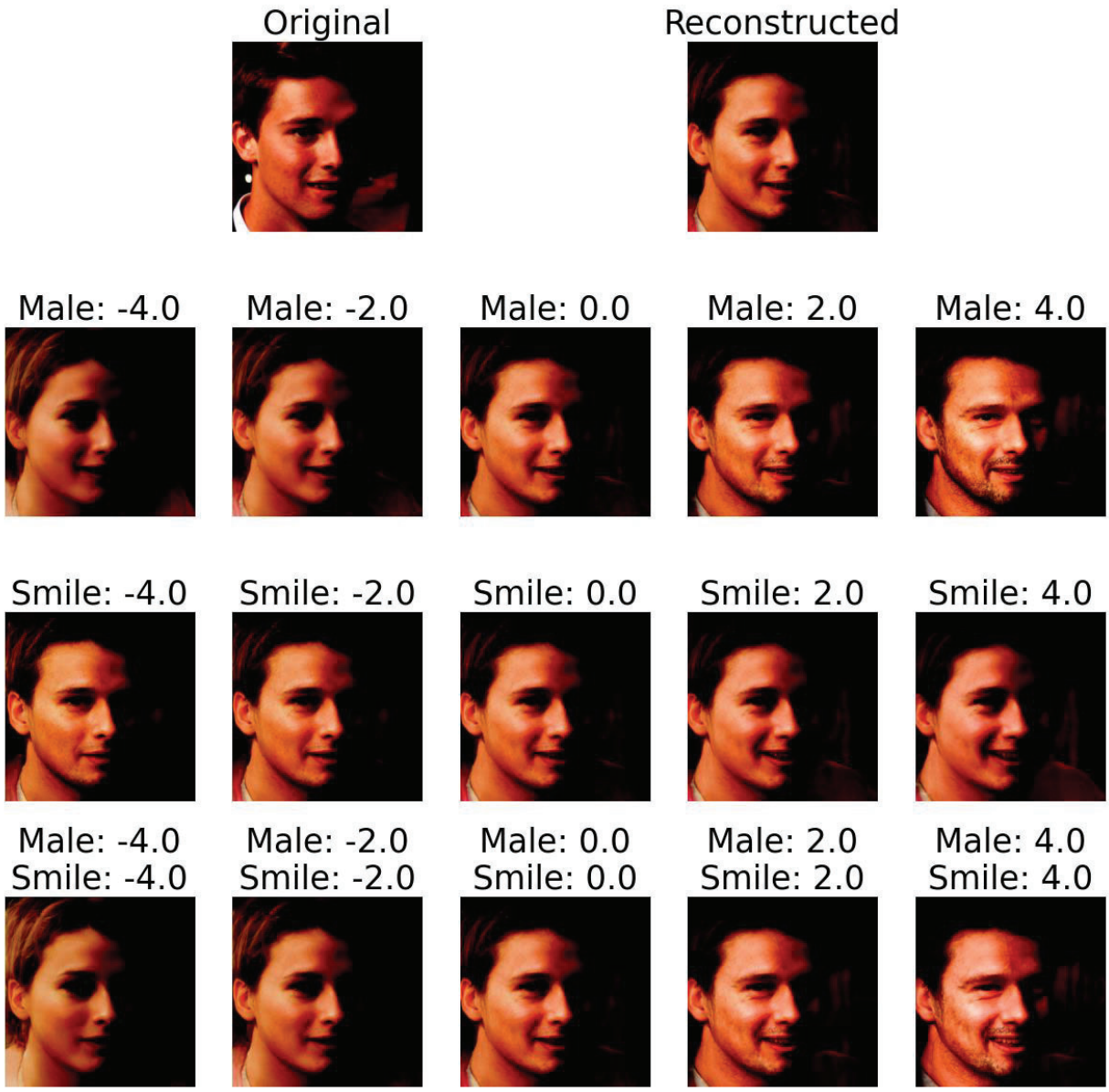

Figure 10: First example output of the GAN based disentanglement network varying the aspects for male/female and smiling/not smiling.

5.3.4 Recreation of attributes

For the purpose of interpreting the distributions of attributes in images of public datasets related to RQ4 the ability of the NICE network to disentangle the w^+ space of an image into $\hat{w}$ space is of importance to explore. Given an image as before we can encode it into the w^+ space and transform it into the $\hat{w}$ space with the NICE network. In this space if the network is able to properly disentangle the class it should be able to find different distributions of values when looking at images on different sides of the binary classification. By separating a dataset into images from one side of the binary classification and another from the other and creating histograms of the values these images have when represented in the $\hat{w}$ space the distributions can be compared. In Figure 11 and Figure 12 the distributions over the classes male and smiling can be found respectively.
The shape of the w^+ space for the pretrained StyleGANv3 model is (16, 512) and due to the bijective transform of the NICE network it needs to maintain the same number of elements resulting in the $\hat{w}$ space also being implemented as the shape of (16, 512). During the training of the NICE network

when creating random vectors in the ŵ space and setting specific values for the controlled attributes one element was assigned the class specific value across the 16 layers. This means that an image that is represented in the ŵ space will have 16 elements representing any one class. For the sake of finding out the distributions of values that represent an image the average of these 16 values are taken.
It is also worth mentioning that while the labels are binary for the images of male and female or smiling and not smiling, in reality what the network is thought to find are the underlying features separating these two pools of labels. It is, therefore, thought to find e.g. the features that are associated with masculinity or femininity from the data available. However, this is a spectrum and there are bound to be overlapping regions in the distributions of such features. The network is looking for patterns to try to separate these distributions from each other where less overlap means that more accurate ways have been found to distinguish the two classes. As can be seen in the figures, while there are overlapping areas the two sides of the binary classes clearly have their own distributions cantered around different values, which is indicative of the network being able to find ways to separate the two.
During the training of the NICE network the encoder loss, classification loss and GAN loss were used together and weighted. During this weighting process the significantly largest weight went to the GAN loss, the second largest to the classification loss and then a significantly smaller encoder loss. This is due to the encoder loss not needing a large weight to find ways of separating the distributions and to reduce instability in the training. Increasing the loss or training for more epochs could both be ways to further improve the shift in distributions.

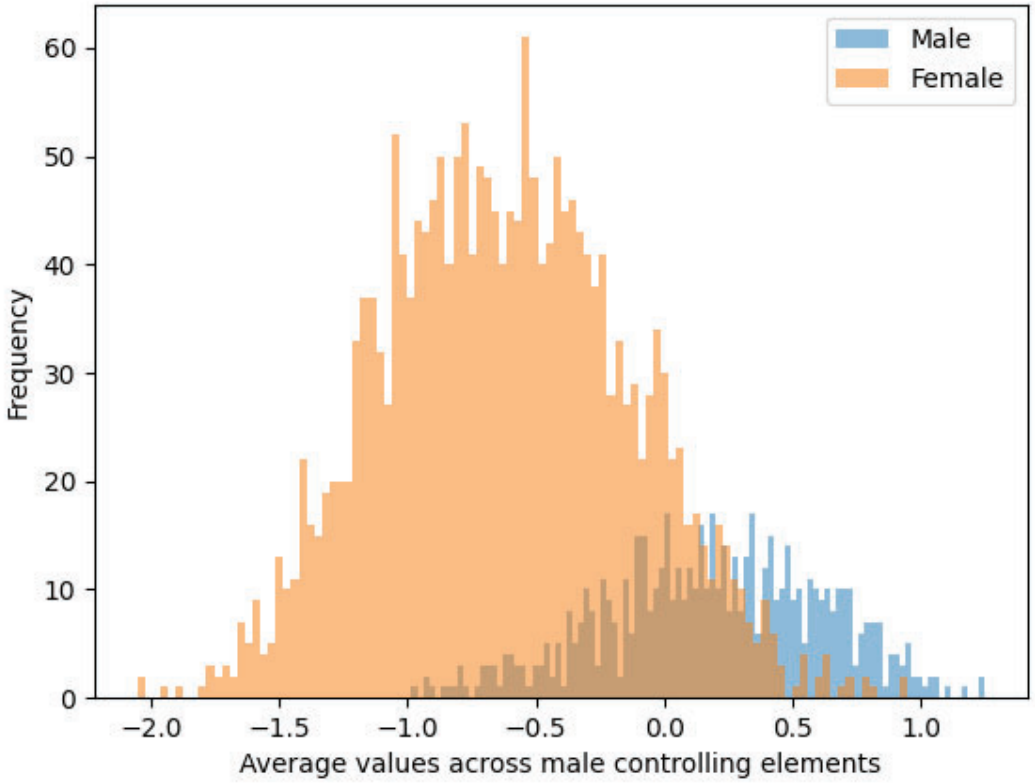

Figure 11: Distribution of averages of male controlling elements in the NICE network for the CelebA – HQ validation dataset for male images and female images.

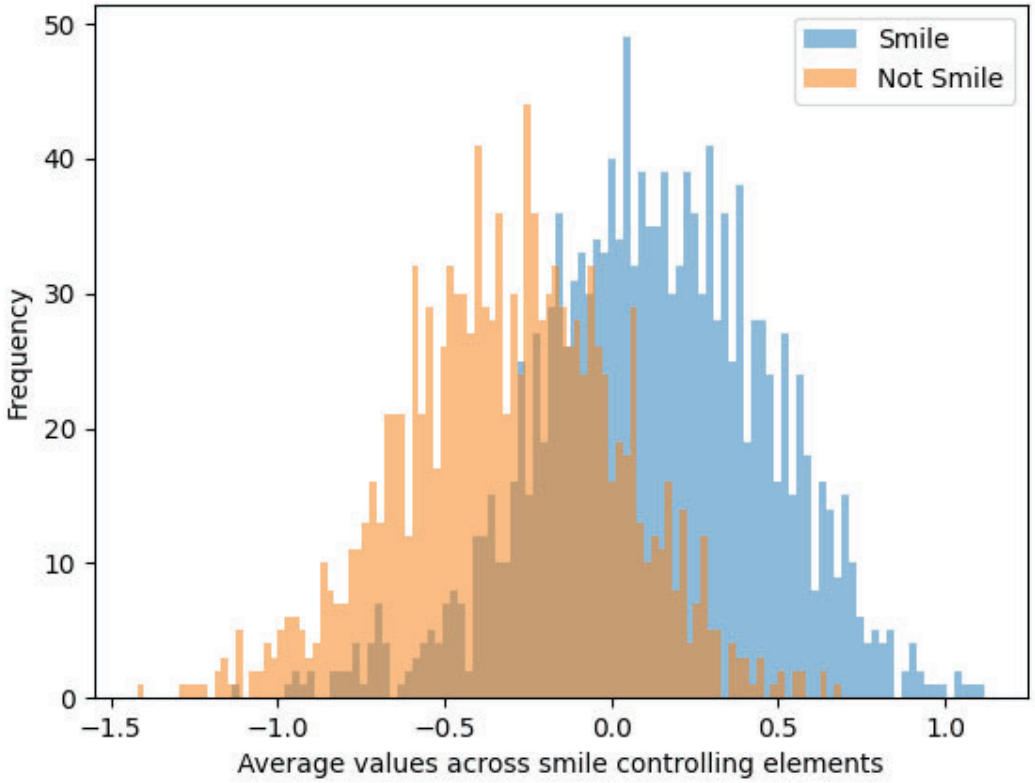

Figure 12: Distribution of averages of smile controlling elements in the NICE network for the CelebA – HQ validation dataset for male images and female images.

6 3DMM-based data generation

6.1 Background

6.1.1 The Basel Face Model (BFM)

Another avenue for the study of disentanglement in face images comes from 3D Morphable Models (3DMM) [22], which have seen significant development, particularly in the realms of disentanglement and novel face generation. Leveraging these models, researchers have developed techniques for disentangled face representation learning, enabling manipulation of individual attributes independently such as identity, expression, pose, and illumination [22].

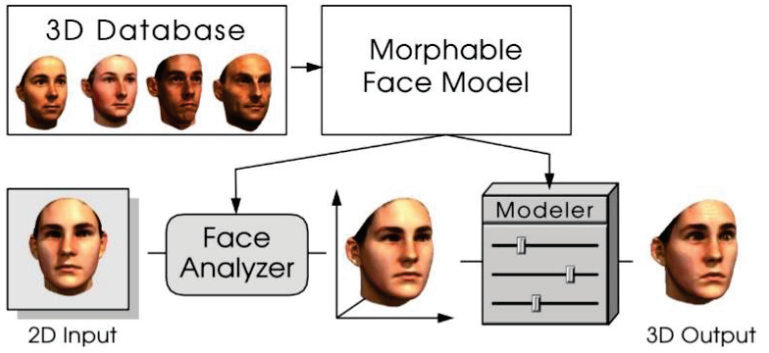

Figure 13: 3DMM face generation pipeline.

One of the pioneering works in the field of 3DMM is the Basel Face Model (BFM) [3], developed by a team of researchers at the University of Basel. The BFM is a statistical model that captures the shape and texture variations of human faces from a large database of 3D scans. It provides a compact representation of facial geometry and appearance, making it suitable for a wide range of applications, and its comprehensive representation of facial shape and texture variations makes it an ideal choice for building statistical models of faces.

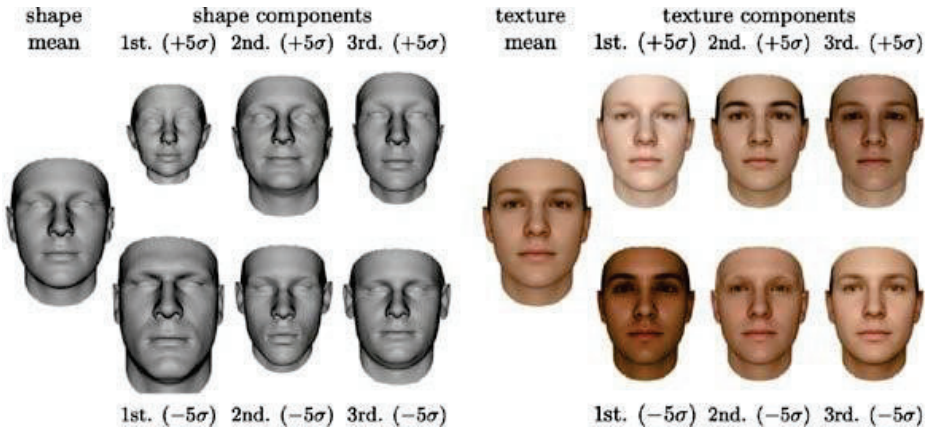

Figure 14: BFM shape and texture models [3]

6.2 Method

6.2.1 Generating NIR facial image data using BFM

All work described in this section makes use of the Basel Face Model (BFM) to generate new faces in 3D. More specifically, we use the first 80 parameters for the BFM shape and texture models and

randomly sample parameters to generate new faces. We investigate various ways of sampling as well as controlling specific features using extracted feature vectors. Once a 3D face has been generated, we make use of computer graphics software to render images of the generated faces.

Importantly, our goal in this section is to generate a synthetic dataset that is beneficial when training a face authentication model. Further, we will use the IR-Face dataset as our real reference dataset, containing images of faces in near-infrared (NIR) and aim to generate synthetic face images that emulate it, but with novel individuals that are demographically diverse in terms of e.g. ethnicity and gender.

6.2.2 Modelling ethnicity in the BFM parameter space (RQ2)

In the original implementation of the Basel Face Model, the authors derived vectors in parameter space representing changes in gender, age, weight, facial height, as well as a separate parameter space for facial expression. These can be used to control and balance the corresponding features. However, their work did not include any vectors representing ethnicity, which is an important feature to balance for a face authentication dataset. Further, the authors acknowledge that the BFM model was primarily based on 3D scans of Caucasian individuals, introducing an inherent bias towards Caucasian facial traits. In order to train a less biased model face authentication model, a balanced dataset representing all ethnicities and genders is required.

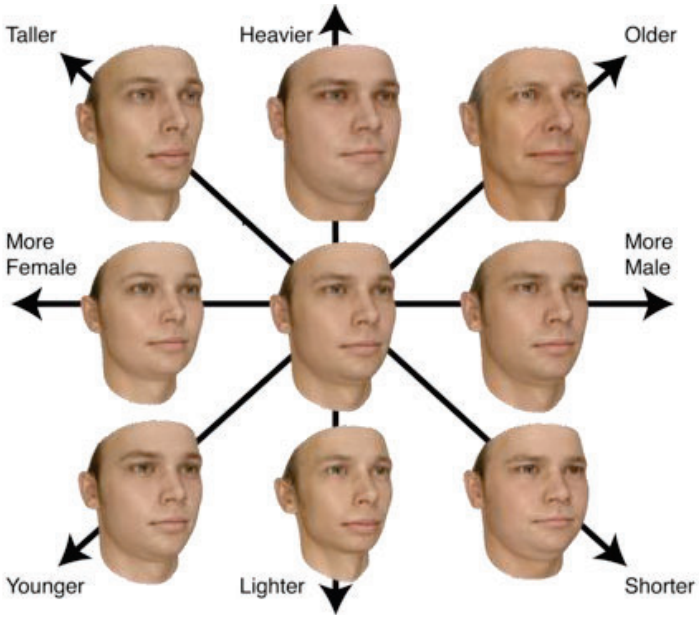

Figure 15: BFM feature vectors. [3]

To allow us to control and balance our synthetic datasets, we therefore investigate deriving vectors in the BFM parameter space representing different ethnicities. Generating synthetic faces of any ethnicity using BFM essentially amounts to modelling their distributions in the shape and texture parameter spaces. In theory, these distributions may be complicated and require lots of data to model accurately. Therefore, we make the simplifying assumption that these distributions can be approximated by a symmetrical distribution, particularly a Gaussian distribution. Generating a given ethnicity therefore only amounts to finding the appropriate parameters for this Gaussian distribution, i.e. its mean and standard deviation, for any ethnicity we want to generate. The mean vector is what we refer to as a "ethnicity vector", as it represents a position in parameter space that when randomized around provides faces of that ethnicity.

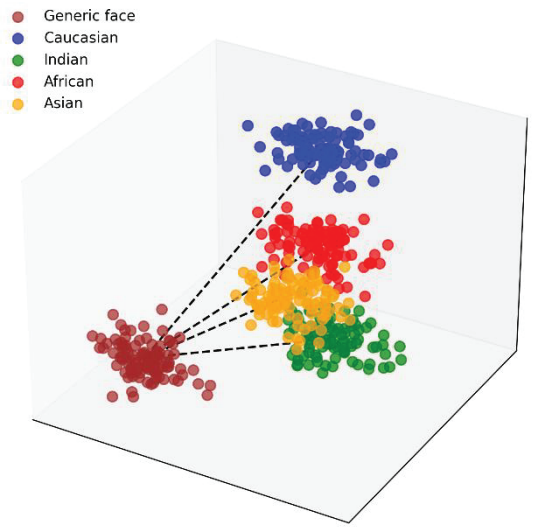

Figure 16: Illustration of different ethnicities modelled as Gaussian distributions in parameter space, here shown in only 3 dimensions. Note that the image is for illustrative purposes only, as the actual BFM parameter shape and texture spaces are 80-dimensional.

To extract these ethnicity vectors, we needed a dataset of real faces with a diversity of ethnicities and the corresponding BFM parameters for each face. No such dataset exists publicly to our knowledge, so our chosen approach is instead to use a CNN model to predict the BFM shape and texture parameters for an existing real face dataset. We use a pre-trained model from [24] for the parameter prediction and FairFace [25] as our real face dataset (see Appendix A for more details). This gives us a large dataset of BFM shape and texture parameters with matching gender, ethnicity and age annotations from FairFace.

Using these, we compute the ethnicity vectors described above as simply the mean of a large number of randomly sampled individuals of each ethnicity. We make sure to balance the sampled individuals with respect to age and gender. We evaluate how representative our ethnicity vectors are by projecting the shape vector of each individual in the FairFace test set onto each of our ethnicity vectors and classify their ethnicity as the projection with the highest component. We compute a confusion matrix for the resulting classifications to show how the predictions correlate with the ground truth labels.

6.2.3 Modelling gender in the BFM parameter space

As described previously, the original BFM included feature vectors for attributes such as gender. Since gender is an important aspect of face authentication datasets, we would like to ensure our generated synthetic datasets are balanced with respect to it. Therefore, we chose to investigate how representative of actual gender the BFM gender vector is.

To do this, we used a similar approach as described in chapter 6.2.2, using a pre-trained CNN from [26] to predict BFM shape parameters for all faces in the FairFace training set. For these experiments, we chose to focus our attention on the shape parameters and ignored the texture parameters. Once extracted, we computed the component of each shape vector in the direction represented by the BFM gender shape vector by simply doing a projection onto it. This gave us a "gender coefficient" for every face in the FairFace dataset, which we could then compare to the corresponding annotated gender.

Our results showed significant overlap between the distributions of males and females from the FairFace dataset with respect to the resulting BFM gender coefficients. To address this issue, we created an improved version of the gender vector that better separates the two genders. We make use of the already predicted shape vectors from FairFace and separate them into two groups for male and female and analyze their distributions in the shape parameter space. To find the vector that best separates the two distributions, we train an SVM to classify gender given the 80 shape parameters of the individuals in the FairFace training set. The resulting model classifies gender using a hyperplane, and the normal vector to this plane should therefore be most representative of gender. To refine the results further, we extracted a separate gender vector for each ethnicity using the same approach, instead of a generalized one. We found this important due to different ethnicities exhibiting distinct prevalent traits, and a single gender vector for all ethnicities fails to capture these nuances.

6.2.4 Data generation pipeline

As mentioned previously, the BFM consists of a shape model, which describes the facial shape as a parameterized set of vertices in 3D space, and a texture model, which describes the color of each vertex [3]. By varying the shape and texture parameters, the BFM can render a large variety of unique faces, providing significant variation between individuals, which is useful for generating a synthetic face recognition dataset.

However, there are a few limitations of the BFM when using it for generating a realistic and diverse facial dataset. Since BFM only models the shape and texture of the head, it lacks the ability to capture some key variations commonly found in real-world facial images, such as hair (both head and facial), accessories like glasses, face masks, or piercings, as well as changes in background or occlusions from hands and other objects.

To address these limitations, we couple the BFM with separate rendering engines that allow us to render the model in a variety of poses and environments as well as with variations in lighting, facial attributes such as facial hair and accessories such as glasses.

6.2.4.1 Rasterizer-based rendering

We began by rendering the BFM using a simple rasterizer in Python using the nvdiffrast [27] package, based on code available from the [26] project.

We generate new unique individuals by randomly sampling the shape and texture vectors of BFM while controlling desired features using the available feature vectors, both from the original BFM model and our ethnicity- and gender-based vectors described above. To simulate varying illumination, we used the spherical harmonics family of functions with randomized coefficients to modify the texture of the face model. When rendering images, we pose the face model in varying poses on top of randomized background images. However, for this approach we did not investigate adding external accessories such as glasses or facial hair, as these require more advanced rendering capabilities.

6.2.4.2 Blender-based rendering

To further improve the realism of the synthetic data created using BFM, we developed a second improved rendering pipeline using Blender [38]. Blender offers advanced computer graphics techniques such as ray-tracing and advanced material properties to better simulate realistic light conditions, as well as support for importing and posing 3D assets together with the BFM.

As described previously, we created new unique individuals by randomly sampling the shape and texture vectors of BFM and used the available feature vectors to control desired attributes such as age, gender and ethnicity. The resulting shape and texture vectors were then used to generate a textured mesh that was imported into Blender.

Once imported into Blender, the face meshes were enhanced with additional elements such as facial hair, glasses, lighting variations, and diverse backgrounds. This was done to provide more variation for each generated individual. Further, we made use of Blender's ray-tracing engine and virtual light sources to simulate realistic lighting. We randomly varied the position and intensity of these light sources. Since our aim is to generate data that matches the IR-Face dataset, we also created custom skin and eye materials in Blender and tuned their light reflecting properties to better simulate the appearance of real faces in near-infrared, such as the specular and diffuse reflectivities.

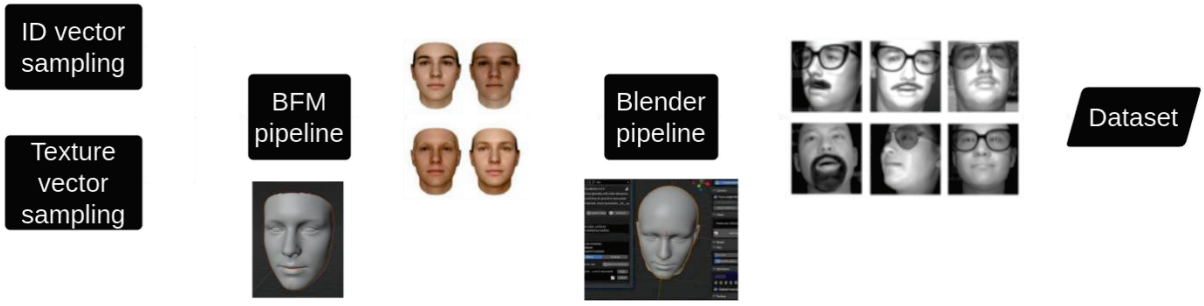

Figure 17: Visualization of the Blender-based rendering pipeline.

6.2.5 Evaluating dataset realism using the Fréchet Inception Distance

To quantitatively evaluate the realism of each rendering method, we employed the Fréchet Inception Distance (FID) score [28], which is commonly used to evaluate the realism of synthetic images by comparing their distribution to real images. The FID score provides a quantitative measure of the similarity between synthetic and real datasets by embedding the images into a feature space using a pre-trained deep neural network, in our case the Inception-v3 network [29], and comparing their statistical properties. A lower FID score indicates higher similarity between the synthetic and real datasets, implying that the synthetic data has achieved greater realism.

Mathematically, FID is defined as:

$$FID(X,Y) = \left|\left|\mu_X - \mu_Y\right|\right|^2 + Tr\left(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{(1/2)}\right)$$

where:

- μ_X and μ_Y are the means of the feature embeddings from the synthetic and real datasets, respectively.
- Σ_X and Σ_Y are the corresponding covariance matrices.
- $Tr(.)$ refers to the trace operation, which sums the diagonal elements of a matrix.

To perform the test, we generated two synthetic datasets: one using the rasterizer-based pipeline and one with the Blender-based pipeline. For the real dataset required in the FID calculation, we utilized the IR-Face dataset (see Appendix A for more details), which consists of face images captured in the Near-Infrared (NIR) spectrum.

6.2.6 Evaluating face authentication performance gains

Since the goal of our generated synthetic data is ultimately to improve face authentication performance, we perform several experiments where we train a face authentication model on the real IR-Face dataset (see Appendix A for more details) as well as our generated synthetic data. Due to showing superior realism based on the FID evaluation, we choose the Blender-based rendering approach for generating the synthetic datasets for the experiments in this section. The face authentication model used in our experiments uses the widely used ArcFace architecture [30].

ArcFace uses a convolutional neural network (CNN) to extract discriminative feature vectors directly from raw facial images and is trained using an angular margin-based loss function. For the CNN backbone, we use the MobileNetV2 architecture due to its effectiveness in learning discriminative features while maintaining computational efficiency [31].

Our experiments use the real IR-Face dataset as a baseline for comparison. The dataset is skewed with respect to gender, containing more male identities than female. Therefore, this provides an opportunity to use the gender vector of chapter 6.2.3 to balance our generated synthetic data to compensate for the imbalance of IR-Face. We hypothesize that a synthetic dataset that is purposefully balanced with respect to gender should provide a better performing face authentication model compared to a synthetic dataset with a gender imbalance.

Therefore, we created two synthetic datasets for our experiments: a balanced dataset and an unbalanced dataset based on the BFM gender distribution. The unbalanced dataset consists solely of male individuals, whereas the balanced dataset includes an equal distribution of male and female individuals. We confirm that the two datasets have the correct gender distributions by analysing the distribution of gender coefficients along our improved gender vector described in chapter 6.2.3.

The face authentication experiments performed are structured as follows:

1. **Baseline face authentication with IR-Face dataset only**:
   The model is trained on the original IR-Face dataset without any balancing. This serves as the baseline for comparison for the following two experiments.

2. **Face authentication with IR-Face dataset + BFM Unbalanced**:
   Here, the model will be trained on the original IR-Face dataset and an unbalaced synthetic dataset that contains only male individuals in order to test the impact of the gender imbalance on the face authentication task.

3. **Face authentication with IR-Face dataset + BFM Balanced:**
   In this experiment, the model will be trained on the original IR-Face dataset and a synthetic dataset that has been balanced using our extracted gender vector, to confirm the hypothesis that gender balancing is important when generating synthetic data for the face authentication task.

Table 2 shows the details regarding all the datasets used in these tests. Detailed descriptions of these datasets can be found in Appendix A. For all experiments, we evaluate the resulting models on the IR-Face test set using False Acceptance Rate (FAR) and False Rejection Rate (FRR).

Dataset name	Total subjects	Male subjects	Female subjects	Number of images per subject	Total number of images
IR-Face	1028	1085	146	200	205 600
BFM Balanced	1000	500	500	200	200 000
BFM Unbalanced	1000	1000	0	200	200 000

Table 2: Details of the datasets used for the face authentication experiments.

6.2.6.1 Evaluation metrics – FAR and FRR

We evaluate our face authentication model using the commonly used metrics False Acceptance Rate (FAR) and False Rejection Rate (FRR). These metrics provide insights into the model's ability to

distinguish between legitimate users and impostors, thus determining its overall security and reliability. The False Acceptance Rate (FAR) measures the probability of the model incorrectly accepting an impostor as a legitimate user, while the False Rejection Rate (FRR) measures the probability of the model incorrectly rejecting the legitimate user. Ideally, a robust face authentication model aims to strike a balance between FAR and FRR, minimizing both rates to achieve optimal performance. However, there is often a trade-off between the two: as one rate decreases, the other typically increases, and vice versa.

6.3 Results

6.3.1 Modelling ethnicity in the BFM parameter space

Figure 18 shows a normalized confusion matrix between the annotated ethnicities in the FairFace test dataset and the predicted ethnicity given by projecting each individual's shape vector onto our newly created ethnicity vectors and taking the largest coefficient as the classified ethnicity. As can be seen, there is a strong correlation between the class predicted using this approach and the true annotated ethnicity, indicating that our ethnicity vectors are indeed representative of each ethnicity. Notably however, the Latino Hispanic ethnicity shows the lowest correlation, indicating that it is not well-represented by its ethnicity vector. Further, there is a notable cross-correlation between the East Asian and Southeast Asian ethnicities, which we argue is to be expected since these share many facial attributes.

Figure 19 shows examples of generated faces using our extracted ethnicity vectors, in this case for the African and East Asian ethnicities as annotated in the FairFace dataset.

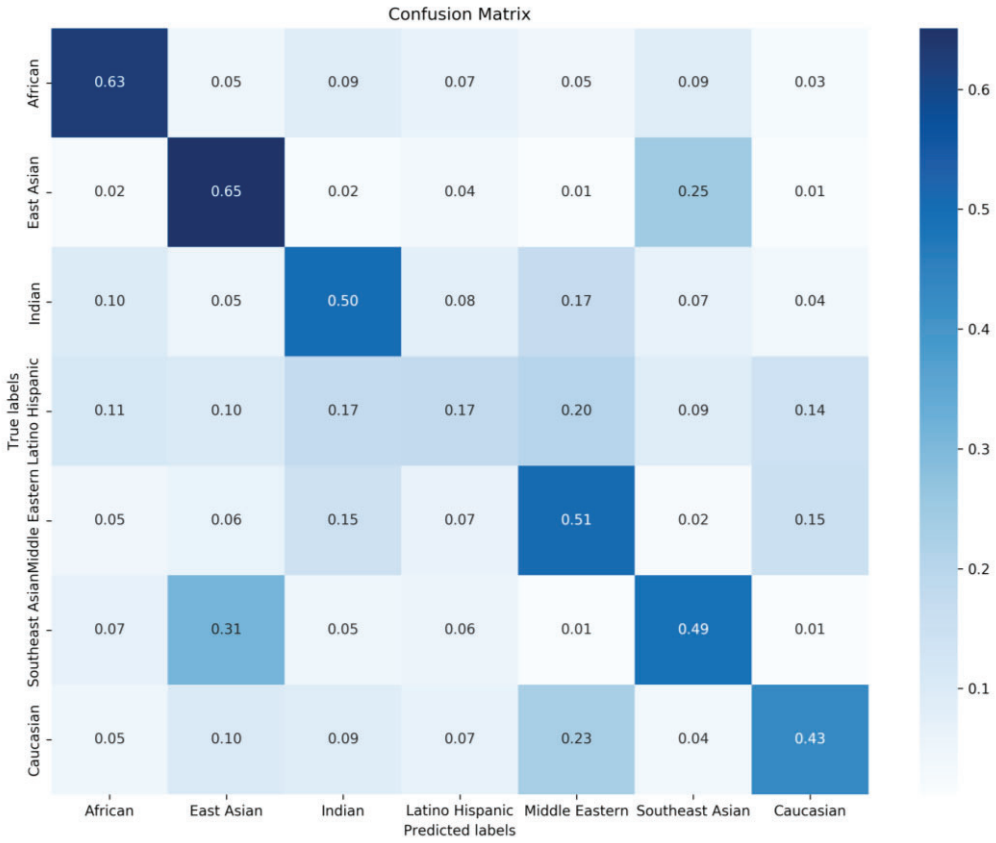

Figure 18: Confusion matrix between the annotated ethnicities in the FairFace test set and the predicted ethnicities based on our extracted ethnicity vectors.

Figure 19: Example of new faces created using the ethnicity vector method.

6.3.2 Modelling gender in the BFM parameter space

Figures 20 and 21 show the resulting gender coefficients for the FairFace dataset when projecting the shape vectors of the identities in the training set onto the original BFM gender vector and the newly created FairFace gender vector. As can be seen, the FairFace gender vector better separates the distributions of males and females, and therefore it can be interpreted as better representing gender.

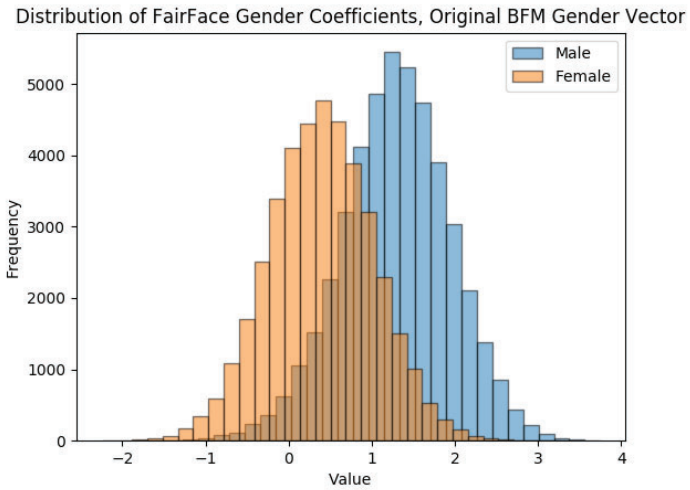

Figure 20: Distribution of FairFace gender coefficients for the original BFM gender vector.

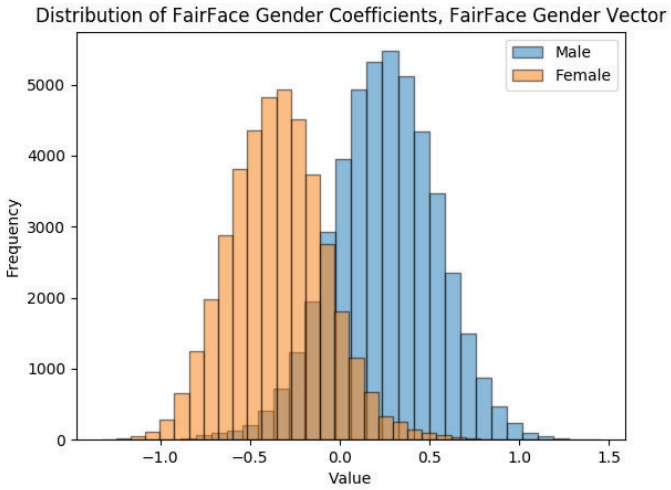

Figure 21: Distribution of FairFace gender coefficients for the newly created FairFace gender vector.

6.3.3 Data generation pipeline

Figure 23 shows example images generated using the rasterizer-based method and figure 24 from the Blender-based method. Visually, we argue that the Blender-based method generates more realistic images, which is also to be expected due to its more advanced rendering capabilities.

More images from the Blender-based method are also shown in Figure 25 to illustrate the high image diversity that can be achieved with this pipeline, varying ethnicity, gender, head pose, lighting, accessories, facial hair, expression and more.

Figure 22: Example real images from the IR-Face dataset.

Figure 23: Example image from the Rasterizer-based pipeline.

Figure 24: Example image from the Blender-based pipeline.

Figure 25: More examples of images created using the Blender-based pipeline, showing the diversity that can be achieved for each individual.

6.3.4 Fréchet Inception Distance realism evaluation results

Table 3 shows the resulting FID scores for the datasets generated using our two rendering methods. Initially, we compared two subsets of the IR-Face dataset to establish a baseline, achieving a score of 15.877. Next, we compared the IR-Face data with the synthetic data generated. The results indicate that the Blender-based pipeline achieved a lower FID score of 71.691, while the rasterizer-based pipeline recorded a score of 93.412. This confirms our visual intuition that the Blender-based pipeline produces images that are more realistic and closer in distribution to the actual data from the IR-Face dataset. It should be noted that neither dataset achieves as low FID as the two real subsets. We hypothesize that this could be due to other differences in their distributions, such as the exact head poses, lighting, crop margins and backgrounds, which also should affect the final FID score.

Datasets	FID score
IR-Face dataset vs IR-Face dataset	15.877
IR-Face dataset vs Rasterizer pipeline	93.412
IR-Face dataset vs Blender pipeline	71.691

Table 3: FID scores between different dataset pairs.

6.3.5 Face authentication experiment results

Figure 26 shows the gender distributions along the FairFace gender vector for the FairFace, IR-Face and our two generated synthetic datasets, BFM Balanced and BFM Unbalanced. The first column shows the full distributions and the second and third columns show the distributions for females and males respectively. As can be seen, the BFM balanced dataset better covers the full range of gender

coefficients and is centered approximately in agreement with the FairFace and IR-Face datasets. This confirms that our gender balancing works as expected.

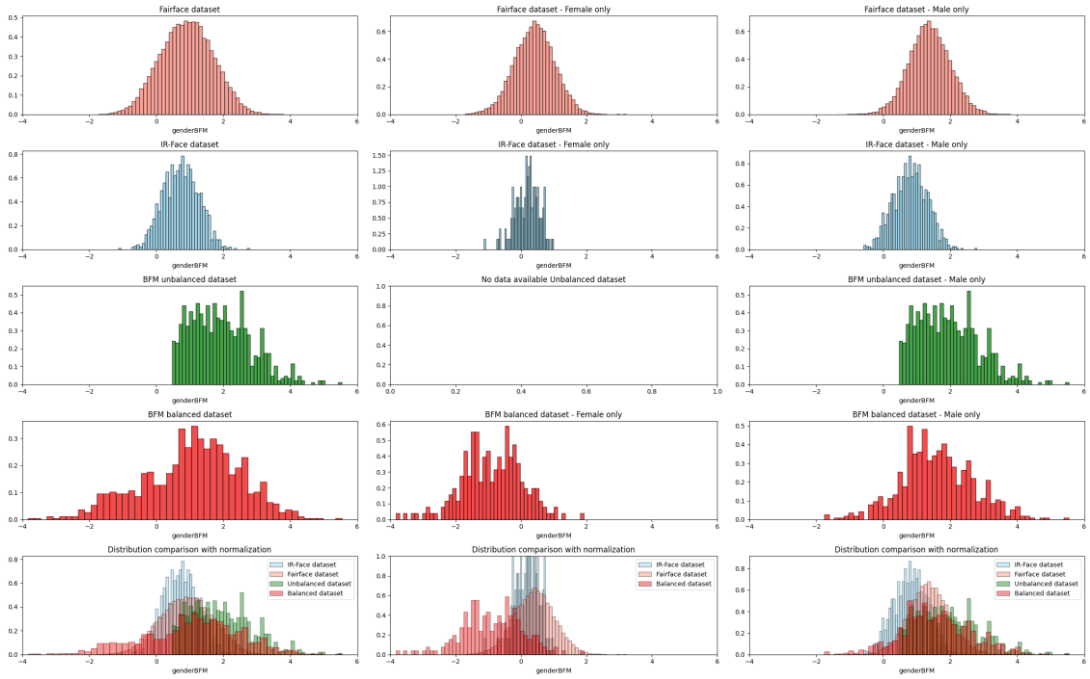

Figure 26: Distribution of the new BFM gender vector coefficients for identities in each dataset.

Table 4 shows the resulting FRR values at the fixed value of FAR= 10^{-5}, evaluated on the IR-Face test dataset. As can be seen, the face authentication model shows an increase in performance for both synthetic datasets compared to the real-only baseline. The best results are achieved IR-Face + BFM Balanced experiment, confirming our hypothesis that a synthetic dataset that is balanced with respect to gender and covers the entire range of gender coefficients is superior to an unbalanced dataset.

Further, Figure 27 shows the ROC curves for the three experiments, where the trade-off between FRR and FAR can be seen. Again, the IR-Face + BFM Balanced training shows the best performance for almost all combinations of FRR and FAR.

Dataset name	Test dataset	FRR@FAR=10^{-5}
IR-Face train dataset	IR-Face test dataset	$10 \cdot 10^{-4}$
IR-Face train dataset + BFM Unbalanced	IR-Face test dataset	$5 \cdot 10^{-4}$
IR-Face train dataset + BFM Balanced	IR-Face test dataset	$1.5 \cdot 10^{-4}$

Table 4: Details of the datasets to use for the face authentication test.

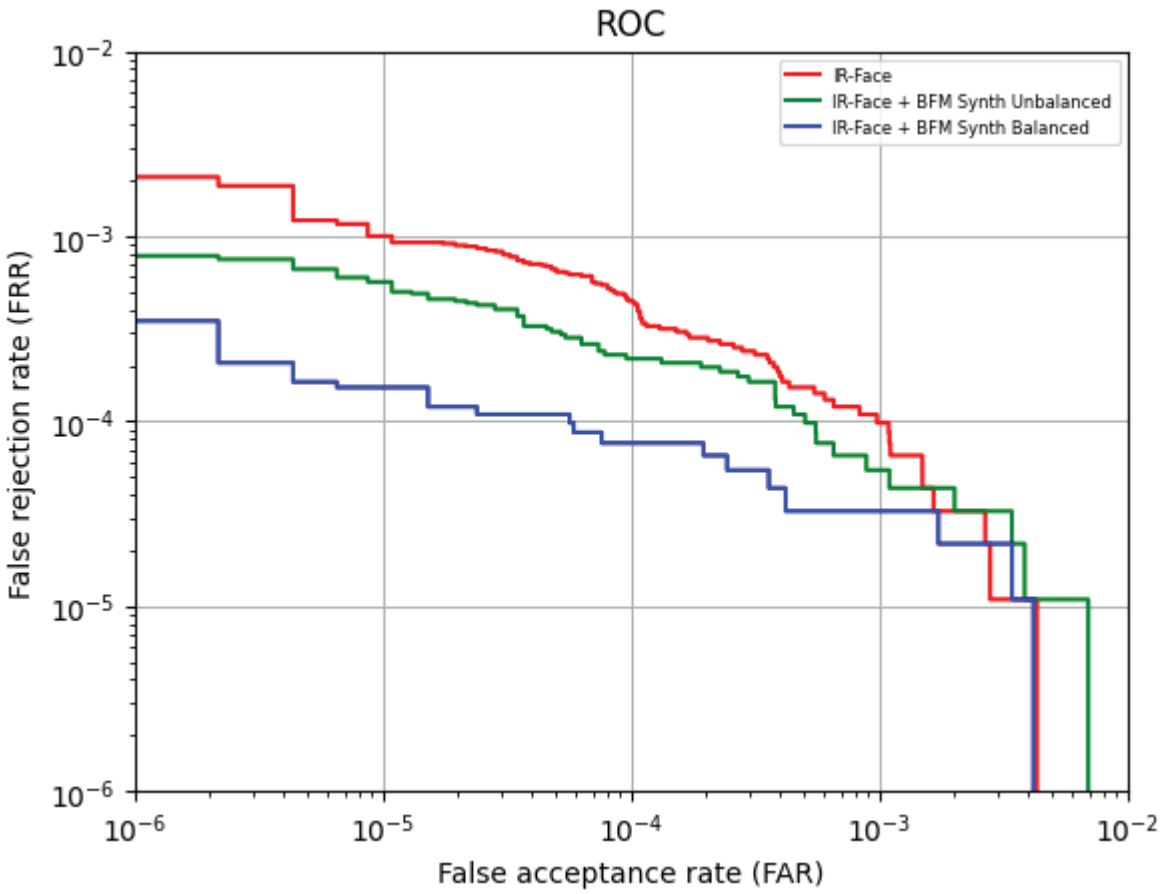

Figure 27: ROC curves for the three models evaluated on the IR-Face test set, showing the trade-off between FRR and FAR.

7 Dissemination and publications

The DIFFUSE project has supported a master thesis student -- Liam Tabibzadeh -- in his studies. He wrote his thesis at RISE within the DIFFUSE project focusing on the structuring of the latent space in the GAN-based approach. In his work, he explored the sorting of features into an identity-related and a non-identity-related half of the latent space. The goal was to create a latent space where two different images of the same individual will have an almost identical representation in the first half of the space, while two otherwise identical photos of different individuals (same background, lighting, perspective, clothing, expression, etc.) will have an almost identical second half of the latent space. Liams thesis has been published and is publicly available at Diva Portal [32].

Internally the partners have been holding meetings and presentations where they are describing the contents and progress of the DIFFUSE project, spreading the information. Information on the DIFFUSE project has been disseminated externally during a PhD disputation, master thesis presentation and through seminars for industrial PhD students from different companies in Sweden as part of a research school.

The DIFFUSE project has been supporting Kateryna Melnyk, employee at RISE at the time, in her PhD thesis dissertation defence where the choice of defence disputation topic was "*On disentangled latent spaces of generative adversarial networks with application to healthcare*" [33].

As per previous communication the PhD student who was intended to do his Licentiate within the frame of the project has decided to stop his studies preventing that deliverable. This in combination with delays in results also means that it has not been possible to complete scientific publications within the project.

The results of this project will be leveraged by Smart Eye to enhance their research and development efforts. By generating data that can be used to enrich their datasets in infrared (IR), the project addresses the challenge of collecting sufficient data to balance existing datasets. This advancement will significantly improve their ability to maintain a demographically well-balanced research dataset for training purposes. Furthermore, it will enable ongoing evaluation of the impact of using generated data in their development processes.

8 Conclusions and future research

The project has been exploring a GAN-based approach for disentanglement. This method includes a strategy for separation of ID and non-ID features of latent spaces from images, that has been explored in a master thesis, for the purpose of improving authentication algorithms. It also covers a method for disentangling a reversible latent space for pretrained generation models. This method allows for training the disentanglement based on annotations in available dataset and has been tested with the CelebA – HQ dataset.
Further, within the GAN-based approach an image classifier has been trained that classifies two of the attributes in CelebA and has achieved an average accuracy on the validation dataset of 96%. The complete GAN-based network has been shown to be capable of editing real images changing specific attributes, such as whether the person is male or female in parallel to smiling or not smiling, while maintaining the general look of the image. The model's capabilities to extract representations of attributes from real images has also been tested showing a clear distinction of the distributions for the binary classes.

Additionally, the project has developed a method for generating synthetic data for face authentication based on the Basel Face Model (BFM). One of the shortcomings of the original BFM method is its limitation in the number of attributes and the fact that it was created using data heavily skewed towards Caucasians. The developed method significantly enhances gender and ethnicity separation, allowing for greater diversity in the data. This approach has been applied to balance datasets and improve face authentication methods.

The face authentication experiments demonstrated a clear improvements in both the False Acceptance Rate (FAR) and False Rejection Rate (FRR) when synthetic data was added. The best results were obtained when the IR-Face dataset was combined with the BFM Balanced dataset, indicating that gender balancing is an important aspect to consider when generating a synthetic dataset.

Furthermore, the FID scores from the two rendering methods support these findings. First, we established a baseline by comparing two subsets of the IR-Face dataset, achieving a score of 15.877. When comparing the IR-Face data with synthetic data, the Blender-based pipeline produced a lower FID score of 71.691, indicating that it generates more realistic images. In contrast, the rasterizer-based pipeline recorded a higher score of 93.412, confirming our visual intuition that the Blender-based method creates images closer in distribution to the actual data from the IR-Face dataset.

8.1 Future research with GAN-based approach

Among the ways the method could be improved in the future is to add an additional loss function to ensure disentanglement. At the current point it enforces specific annotations to be controlled by specific elements, but it does not actively discourage other elements. Given that there would be enough attributes used in the training and each are encouraged into its own element the current method could potentially accomplish this. As it is now there is nothing saying that the same element that is controlling smiling or not also controls hair colour. As discussed earlier if hair colour had its own element, then the information of it would naturally be pushed into that element instead of others like that of smile. Thus, it could be beneficial to increase the number of annotations used both for the utility and also for the disentanglement factor. Alternatively, to disentangle other elements not seen before an additional loss that discourage other concepts rather than the one assigned to the element could be beneficial to research.

With multiple losses for both the T network and for the discriminator in addition to two separate stages of training there is a need for balancing. If one of the loss functions becomes dominant it is likely that the others will not be given as much thought and can even cause the GAN like training to collapse. Finding a good balance between the different losses requires many tries and finetuning. For the sake of the tests in this project the losses have been set to be more in favour of the discriminator loss than the classification loss, but it could be possible to increase the gap further which enforces better realism in images at the cost of separations of the disentangled latent space. The idea is that while it becomes clear at a proof-of-concept stage that the method is working, more finetuning and longer training sessions can be added at a later stage to improve upon image quality.

8.2 Future research with 3DMM-based approach

Key findings from this research highlight the potential of synthetic data to improve both diversity and performance in face authentication tasks. Future research will focus on making synthetic data even more realistic by enhancing control over features such as age, facial expressions, and lighting conditions. Additionally, efforts will be made to achieve a clearer separation between ethnicities and gender in the vector space, ensuring more accurate and distinct representation of these attributes. Further work will also aim to extend the Blender-based generation pipeline to include more type of accessories and occlusions by other objects to generate data that more closely resembles real datasets. Ultimately, these advancements could enable more robust face recognition systems that are less biased and perform better across various demographic groups, thereby improving fairness and inclusivity in facial recognition technology.

9 Participating parties and contact persons

RISE Research Institutes of Sweden AB

Lindholmspiren 3A

417 56 Göteborg

Contact: Martin Torstensson

Martin.torstensson@ri.se

Smart Eye Aktiebolag

Masthamnsgatan 3

413 27 Göteborg

Contact: Henrik Lind

Henrik.lind@smarteye.se

Högskolan i Halmstad

Kristian IV:s väg 3

301 18 Halmstad

Contact: Fernando Alonso-Fernandez

fernando.alonso-fernandez@hh.se

10 References

[1] "DRAMA-2 Driver and passenger Activity Mapping simulator", published by RISE Research Institutes of Sweden AB, https://www.ri.se/en/what-we-do/projects/drama-2-driver-and-passenger-activity-mapping-simulator (accessed: 02-10-2024)

[2] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. In Proc. NeurIPS.

[3] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In Proceedings of the 6th IEEE international conference on advanced video and signal based surveillance (pp. 296-301).

[4] I. J. Goodfellow et al., "Generative Adversarial Networks," arXiv:1406.2661 [cs, stat], Jun. 2014, Accessed: Feb. 26, 2021. [Online]. Available: http://arxiv.org/abs/1406.2661

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114 [cs, stat], May 2014, Accessed: Dec. 06, 2021. [Online]. Available: http://arxiv.org/abs/1312.6114

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," CoRR, vol. abs/1710.10196, 2017, [Online]. Available: http://arxiv.org/abs/1710.10196

[7] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," CoRR, vol. abs/1812.04948, 2018, [Online]. Available: http://arxiv.org/abs/1812.04948

[8] Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, pp. 8107–8116. doi: 10.1109/CVPR42600.2020.00813.

[9] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[10] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data," CoRR, vol. abs/2006.06676, 2020, [Online]. Available: https://arxiv.org/abs/2006.06676

[11] Zhu, J., Zhao, D., Zhang, B., and Zhou, B. (2022). Disentangled infer- ence for gans with latently invertible autoencoder. International Journal of Computer Vision, 130(5):1259–1276.

[12] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets," in Advances in Neural Information Processing Systems, 2016, vol. 29. Accessed: Sep. 21, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html

[13] W. Nie et al., "Semi-Supervised StyleGAN for Disentanglement Learning," in Proceedings of the 37th International Conference on Machine Learning, Nov. 2020, pp. 7360–7369. Accessed: Dec. 06, 2021. [Online]. Available: https://proceedings.mlr.press/v119/nie20a.html

[14] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh, "InfoGAN-CR and ModelCentrality: Self-supervised Model Training and Selection for Disentangling GANs," in Proceedings of the 37th

International Conference on Machine Learning, Jul. 2020, vol. 119, pp. 6127–6139. [Online]. Available: https://proceedings.mlr.press/v119/lin20e.html

[15] L. Pan, P. Tang, Z. Chen, and Z. Xu, "Contrastive Disentanglement in Generative Adversarial Networks," CoRR, vol. abs/2103.03636, 2021, [Online]. Available: https://arxiv.org/abs/2103.03636

[16] Wu, Z., Lischinski, D., and Shechtman, E. (2021). Stylespace analysis: Disen- tangled controls for stylegan image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12863–12872.

[17] Shen, Y. and Zhou, B. (2021). Closed-form factorization of latent semantics in gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1532–1540

[18] Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., and Yang, M.-H. (2022). Gan inversion: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):3121–3138.

[19] Dinh, L., Krueger, D., & Bengio, Y. (2014). Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516.

[20] Alaluf, Y., Patashnik, O., Wu, Z., Zamir, A., Shechtman, E., Lischinski, D., and Cohen-Or, D. (2022). Third time's the charm? image and video editing with stylegan3.

[21] Tabibzadeh, L. (2024). Disentangled Latent Spaces for Synthetic Data (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-535764

[22] Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 187-194.

[23] Egger, B., & Tewari, A. (2020). Disentangled Representation Learning with Structured Generative Models: A Review. arXiv preprint arXiv:2011.13859.

[24] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2020). Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. https://arxiv.org/abs/1903.08527

[25] Kärkkäinen, K., & Joo, J. (2019). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. https://arxiv.org/abs/1908.04913

[26] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2020). Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. https://arxiv.org/abs/1903.08527

[27] Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., & Aila, T. (2020). Modular Primitives for High-Performance Differentiable Rendering. https://arxiv.org/abs/2011.03277

[28] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. https://arxiv.org/abs/1706.08500

[29] O. Lang et al., "Explaining in Style: Training a GAN to explain a classifier in StyleSpace," CoRR, vol. abs/2104.13369, 2021, [Online]. Available: https://arxiv.org/abs/2104.13369

[30] Deng, J., Guo, J., Niannan, X., Zafeiriou, S., Trigeorgis, G., & Liu, Z. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690-4699.

[31] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510-4520.

[32] L. Tabibzadeh, "Disentangled Latent Spaces for Synthetic Data," Uppsala University, Department of Information Technology, Master's thesis, 2024. [Online]. Available: https://uu.diva-portal.org/smash/record.jsf?dswid=1187&pid=diva2%3A1887438

[33] K. Melnyk, "On disentangled latent spaces of generative adversarial networks with application to healthcare," Disputation, Freie Universität Berlin, 2024. [Online]. Available: https://www.mi.fu-berlin.de/fb/dates/disputationen/Disputation-Kateryna-Melnyk.html

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang, "Large-scale CelebFaces Attributes (CelebA) Dataset," https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[35] T. Karras, A. Aila, S. Laine, and T. Aila, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10196

[36] Kim, H., et al. "Fairness-aware Face Recognition: Mitigating Gender and Racial Bias via Fair Feature Learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[37] Rudd, Ethan & Günther, Manuel & Boult, Terrance. (2016). MOON: A Mixed Objective Optimization Network for the Recognition of Facial Attributes. 9909. 10.1007/978-3-319-46454-1_2.

[38] Community, B.O., 2018. Blender - a 3D modelling and rendering package, Stichting Blender Foundation, Amsterdam. Available at: http://www.blender.org.

11 Appendix A

Below is a description of datasets both publicly available and ones created within the project contributing to the state-of-art descriptions of datasets in WP2.

11.1 CelebA and CelebA - HQ

Two of the datasets used in this study are the CelebA and CelebA – HQ. The CelebA dataset which contains over 200,000 celebrity images, covering a diverse range of identities, poses, lighting conditions, and backgrounds. These images were collected from various sources, including celebrity websites and social media platforms. Each image is annotated with 40 binary attributes, such as "smiling," "wearing glasses," and "wearing a hat," providing valuable information for attribute prediction tasks[34]. Examples of the dataset can be found in Figure 27.

Figure 27: Example of the CelebA dataset from https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

The CelebA – HQ dataset is a subset of the original CelebA dataset. It contains 30,000 facial images with a fixed resolution of 1024x1024 pixels. Similarly to CelebA it has 40 binary attribute annotations and multiple instances of recurring identities [34], [35].

All annotations are given as binaries, which can be problematic as many of the annotations are in fact on a gradual scale e.g. there can be varying levels of smiles or blurriness. On another note, there are also annotations subjective to the people doing the annotations such as whether a person is attractive. It is, therefore, of importance when using the dataset to consider what of the annotations are in fact helpful for the training of the models.

Despite its popularity and utility, the CelebA dataset is not without its challenges and limitations. One notable limitation is its focus on celebrity images, which may not always represent the diversity present in the general population. Additionally, while the dataset provides annotations for a wide range of attributes, the quality and accuracy of these annotations can vary, potentially impacting the performance of algorithms trained on the dataset. One of the well-known issues with the CelebA dataset is that the classes are in many cases heavily skewed. It does make sense as there are 40 annotations that are either on or off for each image and it is very hard to balance a dataset to be well distributed across all of these. As an example, if the dataset is supposed to be perfectly split 50%/50% around the attributes “Male”, “Mustache” and ”Bags_under_eyes” it would mean that 50% of all images are males, 50% of all images have a moustache and 50% of all images have bags under the

eyes. If an assumption is made that moustache is a rare class amongst women that would mean that all men in the images would need to have a moustache. For a few annotations this would be possible although hard to maintain, with more classes it becomes practically impossible to maintain. In Figure 28 the distribution of the classes can be seen.

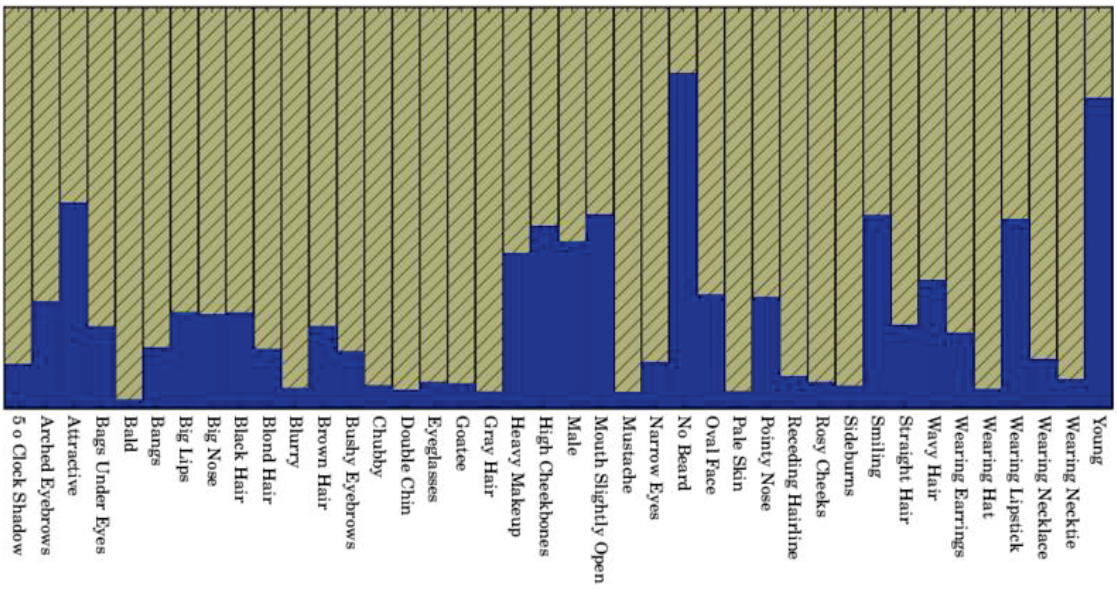

Figure 28: The 40 annotations in the CelebA dataset and their distributions (image courtesy [37]).

11.2 FairFace

The FairFace dataset is a large and diverse collection of facial images specifically designed to address issues of bias and fairness in facial recognition systems. The dataset comprises over 100,000 annotated face images. Each image labelled with attributes including age, gender, and race, covering a wide spectrum of demographic groups. The FairFace dataset is its intentional design to ensure balanced representation across different races and ethnicities, which includes White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino.

Researchers have leveraged the FairFace dataset for tasks ranging from developing fairer facial recognition models to evaluating the performance of existing algorithms on a more demographically balanced dataset. By promoting fairness and inclusivity, the FairFace dataset plays a crucial role in advancing the development of unbiased facial recognition technologies and contributing to the broader goal of ethical AI development [36].

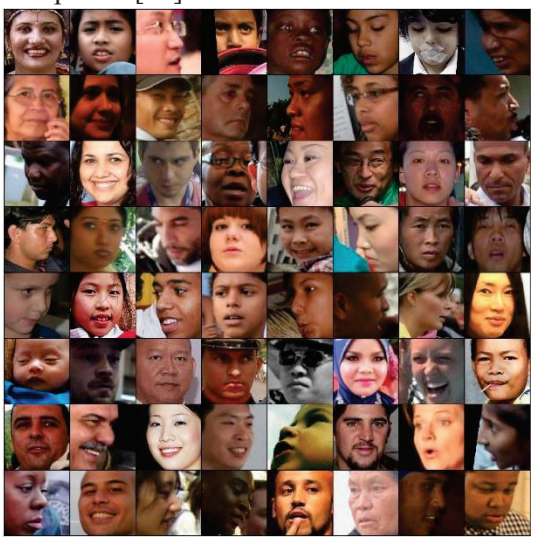

Figure 29: Example of the Fairface dataset from [25].

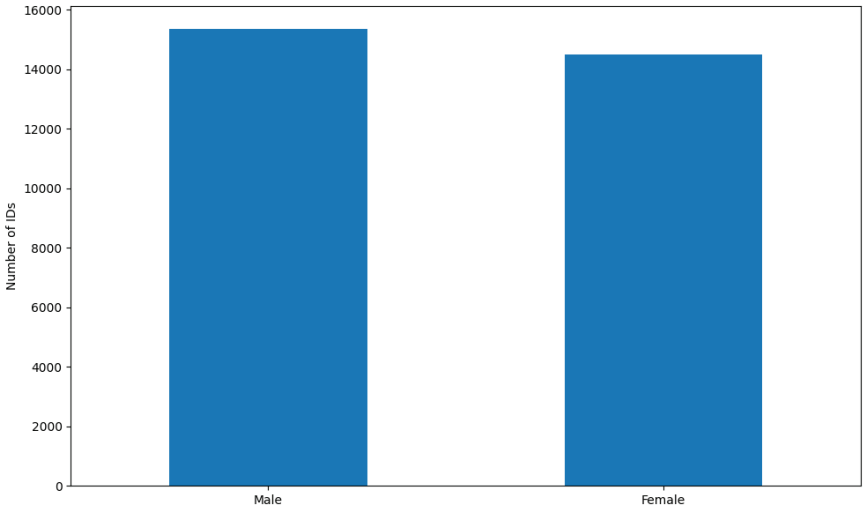

Figure 30: Gender distribution in the FairFace dataset.

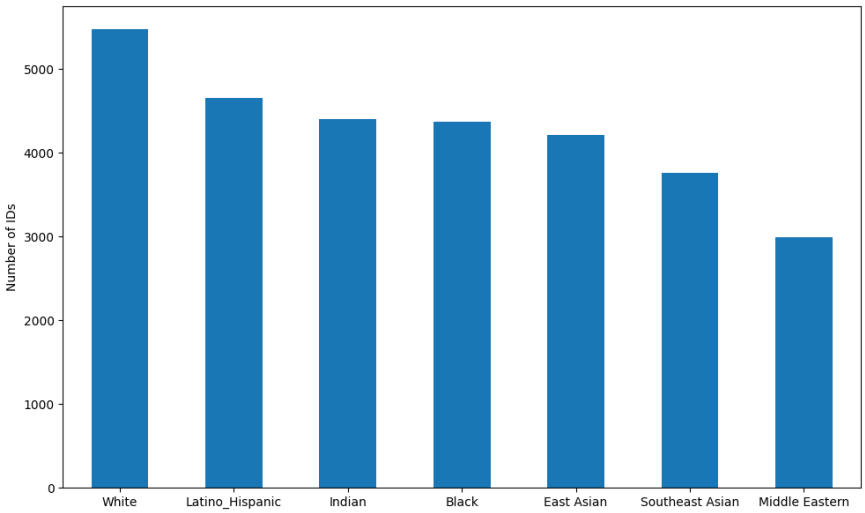

Figure 31: Ethnicity distribution in the FairFace dataset.

11.3 IR-Face

The IR-Face dataset is a collection of facial images captured in Near Infrared (NIR) format. The dataset consists of 1231 unique person IDs, with subjects categorized into five distinct groups: African, Caucasian, East Asian, Middle Eastern, and Other. The Ethnicity Distribution shows a higher prevalence of Caucasian ethnicity, followed by East Asian, Middle Eastern, and African categories. The dataset is split in a way that attempts to balance the ethnicities between the training, validation, and test splits, but there are some imbalances in gender distribution within the dataset, which is particularly pronounced for certain ethnicities with fewer examples.

11.4 BFM Unbalanced

This dataset is generated using the BFM + Blender pipeline and consists of 1000 male identities balanced over ethnicity and age. Each identity is represented by 200 images, showcasing variations in

head pose, lighting, beard styles, accessories such as glasses, and diverse backgrounds to enhance variability.

11.5 BFM Balanced

This dataset is generated using the BFM + Blender pipeline and consists of 1,000 identities, balanced by gender, ethnicity and age, with 500 male and 500 female individuals. Each identity is represented by 200 images, showcasing variations in head pose, lighting, accessories such as glasses, and diverse backgrounds to enhance variability. Additionally, male individuals feature a variety of beard styles and are randomly sampled for the previous BFM Unbalanced dataset.