Toward responsible face datasets: modeling the distribution of a disentangled latent space for sampling face images from demographic groups

Parsa Rahimi[1,2], Christophe Ecabert[1], Sébastien Marcel[1,3]

[1]Idiap Research Institute, [2]EPFL, [3]UNIL

parsa.rahiminoshanagh@epfl.ch, christophe.ecabert@idiap.ch, sebastien.marcel@idiap.ch

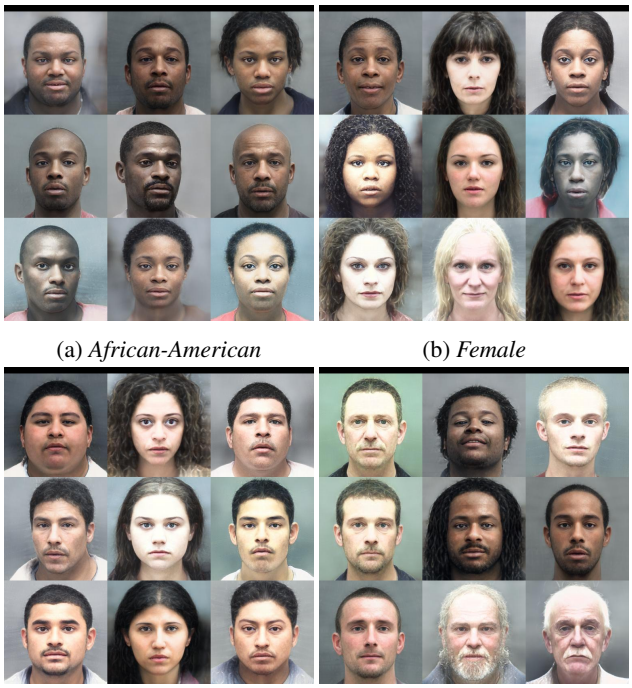
Abstract

Recently, it has been exposed that some modern facial recognition systems could discriminate specific demographic groups and may lead to unfair attention with respect to various facial attributes such as gender and origin. The main reason are the biases inside datasets, unbalanced demographics, used to train theses models. Unfortunately, collecting a large-scale balanced dataset with respect to various demographics is impracticable. In this paper, we investigate as an alternative the generation of a balanced and possibly bias-free synthetic dataset that could be used to train, to regularize or to evaluate deep learning-based facial recognition models. We propose to use a simple method for modeling and sampling a disentangled projection of a StyleGAN latent space to generate any combination of demographic groups (e.g. $hispanic - female$). Our experiments show that we can synthesis any combination of demographic groups effectively and the identities are different from the original training dataset. We also released the source code [1].

1. Introduction

The use of face recognition (FR) systems in critical applications such as law enforcement and recruitment has raised significant ethical concerns. Recent studies have demonstrated that commercially available FR systems using Artificial Intelligence (AI) can exhibit unfairness and bias, particularly against certain demographic groups [6, 52]. As AI systems become more widely adopted in our daily lives, addressing these ethical and legal considerations becomes even more important.

In many cases, the use of FR technology is subject to legal restrictions and regulations, further highlighting the importance of developing fair and accurate systems. One of the main challenges in achieving fairness in FR is the lack of diverse training data, especially considering that the key reason for the success of recent large FR networks is the large datasets which they are being trained upon. At the same time, due to legal and ethical grounds, most of the widely used FR datasets like MS-Celeb1M [20], VGGFace2 [7] and MegaFace [33] have been retracted. Also considering legal policies such as [42, 34], usage of existing datasets including WebFace260M [23] and CASIA-WebFace [18] might also become troublesome when they are deployed in critical applications. Besides these concerns, collecting large amounts of samples required to train



(a) *African-American* (b) *Female*

(c) *Hispanic* (d) *Male*

Figure 1: Generated face images according to desired demographic groups, each 3x3 tile shows images sampled from different demographic groups.

[1] https://gitlab.idiap.ch/biometric/sg_latent_modeling

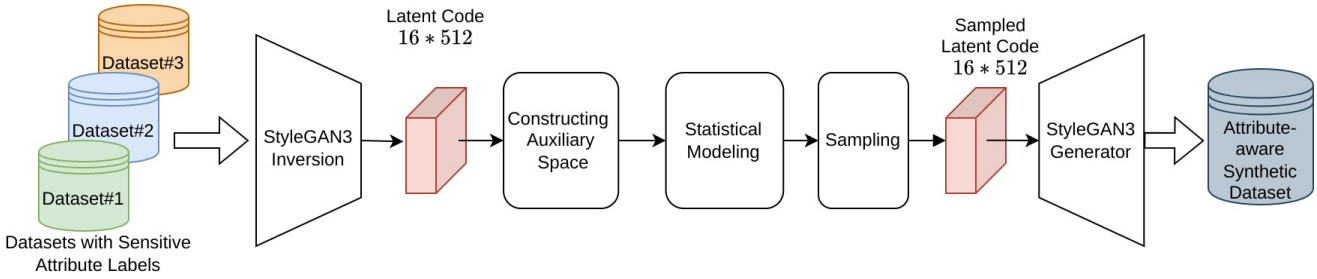


Figure 2: Overall pipeline of our proposed method. Starting from datasets with demogrpahic labels. Usin StyleGAN inversion, we invert the images to desired latent space. To facilitate modeling of StyleGAN latent space we build disentangled auxilarity space. We sample the modeled space to generate desired demogrphic.

deep FR models with various balanced demographic groups is another problem. Therefore, developing complementary datasets that accurately represent underrepresented groups is crucial in mitigating these issues.

The purpose of this work is the creation of balanced face datasets to reduce the bias of the FR models. Currently, approaches for bias mitigation in FR include *pre-processing*, *in-processing*, and *post-processing*. Pre-processing approaches involve modifying the input data to remove or reduce the effects of bias [26, 41, 4]. In-processing approaches work by changing the model architecture or learning algorithm to make it more robust to bias [37, 38, 59, 58, 35, 9, 46]. However, this can compromise fairness, model performance and can be computationally expensive. Post-processing approaches involve adjusting model predictions after training to make them more fair [27, 21, 44, 11, 2], but it can also create a trade-off between fairness and model performance and can be limited by model interpretability.

To address the lack of diversity in existing FR datasets, recent studies propose the use of synthetic data to reduce bias and improve accuracy [55, 56]. However, many of these approaches rely on randomly sampling the latent space of generator models and later attempting to steer and edit the generated signal to meet desired demographics [13]. This can result in accumulation of errors and further biases as the generation is not initially aware of demographic groups. To overcome this limitation and to address the lack of diversity in existing FR datasets, in this paper, we propose a novel yet simple approach to generate such a complementary dataset for FR systems. Fig. 1 show synthetic examples generated by our proposed method. Our generation rely on StyleGAN-based [30, 32] models. This is mostly due to privacy concerns regarding diffusion-based generation. Indeed it was shown in [8] that training data can be inferred from diffusion models which is a limitation for our application scenario to generate new face images.

The proposed method can be expanded to any latent space-based generation architecture. One can see our method as the first step of any demographic editing methods. As we sample desired demographic groups equally, and later on we can employ editing methods like [10, 55, 36, 13] to further generate different variations of same identity to introduce even larger fair datasets.

In Sec. 2 we present related works in the domain of controlled generation and editing of face images. In Sec. 3 we present our approach for controlled face generation. Finally in Sec. 4 we validate our proposed method by various face-related tasks (e.g., demographic classification and identity experiment).

2. Related Works

This section focuses on related works in controlled generation. Additionally, we provide a brief introduction to StyleGAN inversion methods in Sec. 2.4. After examining these methods, it becomes clear that not all of them are suitable for our particular needs.

2.1. Prompt-based Synthesis Methods

Recent advances in generative models especially in diffusion based synthesis [51] and their ability to convert text to often realistic images brought new ways of exploration of generative models. As mentioned previously these methods are often pruned to privacy concerns and also exhibit uncontrollable output.

Using off-the-shelf models (e.g. FairFace classifier [12] and text-image encoders [45]), [62] modeled any control (text-based using CLIP, classifier-based using FairFace classifier) via an energy-based model and try to minimize the divergence between the condition and the supervision of the auxiliary models. By introducing momentum constraint, authors in [62] represented a debiased version of an arbitrary generator.

2.2. Latent-Modeling Methods

Authors in [53] suggested an autoencoder using normalizing flows [15] to form an auxiliary linear separable space. Later one can sample the new space and generate desired demographic groups.

[61] first randomly sampled the latent space of a StyleGAN generator and used an attribute classifier to cluster the input space of StyleGAN. This is done based on the probabilities of the classifier. Finally, using the clustered vectors (prototype vector), authors generate images with the desired attributes.

In this work, by employing an autoencoder with a contrastive loss applied to its bottleneck-layer, we were able to model the complex latent space of any StyleGAN generator with a much simpler modeling technique.

2.3. 3D rendering methods

Recent advances in computer graphics caused the raise of realistic rendering methods that we often see in the gaming and movie industries. Unfortunately, most of these technologies, such as [25] and [24], can not be used because of legal restrictions even for research purposes. However, there are some recent works that generate synthetic datasets using 3D rendering pipelines [3, 64], but they are not as realistic as their commercial counterparts. One benefit of 3D rendering methods is the access to the exact manifold of the models (faces in our case) thus we could easily generate variations of the same identity. As a disadvantage, it is complex to control demographics (e.g., ethnicity) in such methods.

Related to synthetic face dataset generation, authors in [5] trained an identity-conditioned StyleGAN2 [32, 29] to alleviate the privacy concerns of current FR datasets.

2.4. StyleGAN Inversion

StyleGAN inversion is the problem of finding the latent code of an arbitrary image, typically within the domain of the trained network. For example, if the StyleGAN network is trained on face images, the task involves finding the latent code that produces the same image when passed through the synthesis network with similar settings. More specifically, given an input image i and a StyleGAN-based generator G, the goal is to find the latent code that can reconstruct the input image as closely as possible. Inversion methods are generally defined by: (i) latent space in which they map the input image, spaces such as $\mathcal{W}$, $\mathcal{W}^+$, $\mathcal{P}$, $\mathcal{S}$ and (ii) the method used to convert the image to the desired space, such as optimization-based or encoder-based methods. For a more detailed survey of GAN inversion, interested readers may refer to [63].

Here we briefly describe the types of StyleGAN inversion methods proposed in the literature, and we show that not all of these methods are suitable for our application in mind.

Optimization-based : Most of the optimization-based inversion methods change the weights of the synthesis network for each image. As one of the popular methods [50] optimizes the weights of the generators for each image to steer it to a more editable part of the latent space. In this case, we can not reliably model the latent space since the synthesis network would be different for each image.

HyperNetwork-based : The benefits of inversion methods such as those found in [60] and [16] largely stem from weight correction to the generator using an auxiliary network called hypernetwork. This correction is performed based on a per-image-basis, meaning that the original image or its weight correction is required at sampling time. This prevents the synthesis of images solely from the latent space of the generator.

Encoder-based : This type of StyleGAN inversion involves using an auxiliary mapping network to convert the input image to the desired latent space (e.g., $\mathcal{W}$, $\mathcal{S}$, $\mathcal{W}^+$, $\mathcal{P}$ spaces). This includes various techniques depending on the architecture and final latent space of auxiliary networks. The two most renowned methods in this category are [49] and [57].

Our key assumption is that the demographics of an image will remain unchanged after inverting it into the desired latent space and reconstructing it using StyleGAN's generator. To verify this assumption, we conducted a qualitative comparison of reconstructed images obtained from the inversion process in the Sec. 4. We conclude that the encoder-based inversion method described in this section is the optimal method for our application. In particular, we used the pixel2style2pixel (pSp) [49] encoder-based inversion method, due to its superior quality compared to [57].

As mentioned previously, our primary research objective is to supplement current datasets with a balanced and fair version. To accomplish this goal, we must also take into account an essential aspect of the various StyleGAN architectures: the distribution of the generated images closely resemble that of the original datasets. Several studies, [30, 19, 17], have investigated the domain-gap issues that arise in the frequency content of generated images produced by different StyleGAN architectures. As it is shown, the StyleGANv3 [30] generation method is less prone to this problem. Thus we conclude employing this method for our synthesis process.

3. Proposed Method

3.1. Problem Setup

Assume that we have an image dataset $\mathcal{D}$ with domain d (e.g. human face images or animal images) depicted by set $\{\mathcal{D}, d\}$ with demographic groups set $\mathcal{A}$. $\mathcal{A}$ can be defined as $\{\mathcal{A}_{gender}, \mathcal{A}_{race}, \mathcal{A}_{age-group}, ...\}$ in which each of them will take some discrete values (e.g. for $\mathcal{A}_{gender}$ this could be $male$ and $female$ and for $\mathcal{A}_{age-group}$ could be children between age 9 to 14 or young adults between age 18

to 30). Given a StyleGAN generative model, $\mathcal{G}$, trained on the same domain d as in $\mathcal{D}$, our goal here is to model the arbitrary sampling spaces of trained StyleGAN model for being able to generate any combination of demographic groups that were presented in $\mathcal{D}$. As an example, for our FR dataset, we want to generate as many synthetic images of *hispanic male* in his *youth (18-30)* as we want. As mentioned previously, by doing so, our goal would be to alleviate the bias introduced due to the disparity of demographics in current face recognition datasets. Here we limit our experiments to the human faces, the same approach also can be used for any $\{\mathcal{D}, d\}$ and StyleGAN generator $\mathcal{G}$ trained on domain d. Fig. 2 illustrates the complete architecture of our generation pipeline for training and inference. Starting from a dataset with demographic, labels such as MORPH [48], UTKFace [54] or FairFace [28] with the images of human faces, we first invert the images using StyleGAN inversion that was trained for $\mathcal{G}$ (described in Sec. 2.4). More specifically given images of $\mathcal{D}$ as $\mathbf{i}$ and inversion network $\mathcal{I}_{\mathcal{G}}$ we compute the inverted latent code, $\mathbf{w}_j$, as:

$$\forall j \in \{1, ..., |\mathcal{D}|\}; \mathbf{w}_j = \mathcal{I}_{\mathcal{G}}(\mathbf{i}_j) \tag{1}$$

Directly modeling the latent space of StyleGANs (e.g., $\mathcal{W}^+$) is impossible because it forms an entangled representation (i.e., latent dimensions do not control a single demographic). Therefore, we form an auxiliary space to disentangle the representation and hence allow for modeling of this new latent space. We build this auxiliary space using the bottleneck layer of an autoencoder. Finally, by sampling the models according to a specific demographic group (e.g., *white-female* or *hispanic-man* in his *30s*) and passing the sampled latent space to our networks, we were able to generate synthetic datasets with any specific attribute.

3.2. Latent Modeling

We first explored the possibility of modeling the $\mathcal{W}^+$ space (i.e. the output of StyleGAN's inversion [49]). However, as reported in [53] and confirmed by our findings in Sec. 4, this latent space is too complex for being able to model it directly using either bijective transforms (i.e. normalizing flows) or other statistical modeling schemes like GMMs. To alleviate this complexity, we employ an autoencoder network. We denote it in Fig. 3. More specifically:

$$\begin{aligned} &\mathbf{b} = E(\mathbf{w}), \\ &\mathbf{w}^* = D(\mathbf{b}) \quad \text{where:} \quad \mathbf{w}^* \simeq \mathbf{w}. \end{aligned} \tag{2}$$

Here E and D are the encoder and decoder parts of the autoencoder respectively, $\mathbf{b}$ is the bottleneck output of the autoencoder (i.e. output of E). To ensure the $\mathbf{w}^* \simeq \mathbf{w}$ we employ an Euclidean loss between the output of D and input of E ($\mathcal{L}_{Reconstruction}$). To enforce the disentanglement of the sensitive demographic groups we employed a contrastive loss applied to the bottleneck layer of autoencoder.

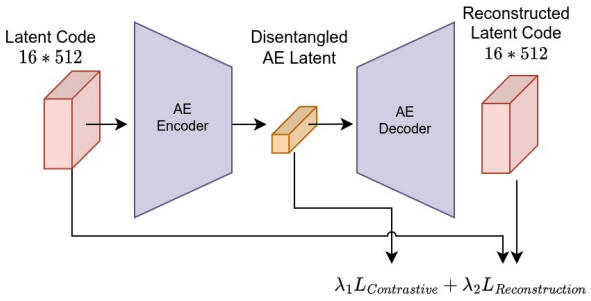


Figure 3: Disentangling latent space using autoencoder with Contrastive Loss

For the contrastive loss, we used the **LiftedStructured** loss proposed in [40] defined as follows:

$$\mathcal{L}_{Contrastive} = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max(0, \mathcal{L}_{i,j})^2 \tag{3}$$

where, $\mathcal{P}$ is the set of positive samples in the mini-batch and the $\mathcal{L}_{i,j}$ is defined as follows:

$$\mathcal{L}_{i,j} = \log\Big(\sum_{(i,k)\in\mathcal{N}} \exp(\alpha - l_{i,k}) + \sum_{(j,l)\in\mathcal{N}} \exp(\alpha - l_{j,l})\Big) + l_{i,j} \tag{4}$$

Here, the function $l_{m,n}$ is a distance function between m-th and n-th samples. We set it as Euclidean distance. $\mathcal{N}$ is the set of negative samples in our mini-batch, and α is the negative margin. By applying contrastive loss on different demographic groups separately, our overall contrastive loss will be the combination of each loss for each demographic group as follows :

$$\mathcal{L}^{Total}_{Contrastive} = \sum_{g\in\mathcal{A}} c_g \mathcal{L}^{g}_{Contrastive} \tag{5}$$

Here, $g \in \mathcal{A}$ means that the contrastive loss is applied to either of $\{\mathcal{A}_{gender}, \mathcal{A}_{race}, \mathcal{A}_{age-group}\}$, separately. In Eq. 5, c_g can be used to control the importance of demographic factors (i.e., $\mathcal{L}^{g}_{Constrastive}$). As mentioned before, for training our autoencoder we also included an Euclidean distance as our reconstruction loss between the $\mathbf{w}$ and $\mathbf{w}^*$, so the total loss will be the weighted sum of reconstruction and contrastive loss as follows:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}^{Total}_{Contrastive} + \lambda_2 \mathcal{L}_{Reconstruction} \tag{6}$$

In Eq. 6, λ_1 and λ_2 are to control the contribution of contrastive and reconstruction loss respectively.

3.3. Gaussian Mixture Modeling

Assuming a disentangled space (i.e. $\mathbf{b}$), we can employ traditional techniques such as Gaussian Mixture Models (GMM) [47] which is defined as follows:

$$\mathcal{M}(\mathbf{b}; \theta_g) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{b}|\mu_m, \mathbf{\Sigma}_m) \tag{7}$$

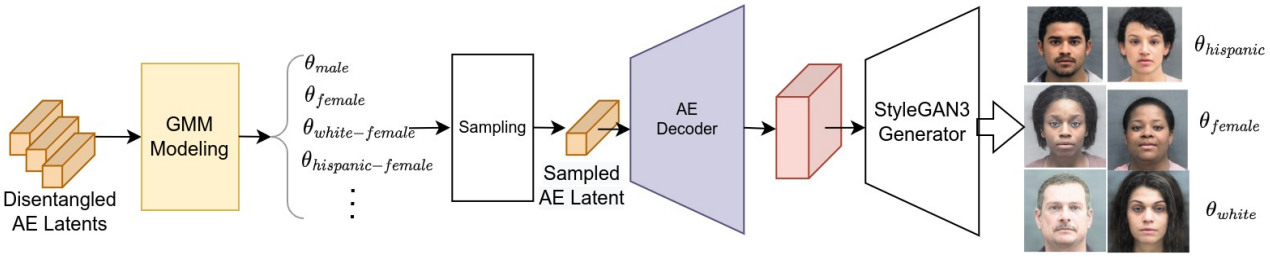


Figure 4: AE-Latent Disentangled Modeling and Sampling

Here, $\mathcal{N}(\mathbf{b}|\mu_m, \mathbf{\Sigma}_m)$ is the multivariate Gaussian distribution, M is the number of mixture components, μ_m, $\mathbf{\Sigma_m}$ and w_m are mean, covariance matrix and weight of mixture component number m respectively. The weights must satisfy $\sum_{m=1}^{M} w_m = 1$. θ_g is a set of all the mentioned parameters. To model the space for a given demographic group, g, we use the Expectation-maximization [39] algorithm on the samples in g demographic group to solve for parameters θ_g. As an example, we fit a GMM to the $male$ group and another one for a $hispanic - female$ demographic, respectively denoted by $\mathcal{M}(\mathbf{b}; \theta_{male})$ and $\mathcal{M}(\mathbf{b}; \theta_{hispanic-female})$. Here we can compute the likelihood of a sample being drawn from g as $P(\mathbf{b}|\theta_g)$, likewise, the log-likelihood (LL) can be formulated as $LL = log(P(\mathbf{b}|\theta_g))$.

3.4. Generating Images with the proposed approach

As shown in Fig. 4, we first sample $\mathbf{b}$ according to desired demographic groups by using their corresponding GMM parameters (i.e., θ_g). Then, we use decoder part of our autoencoder ($D(\mathbf{b})$) to obtain latent code that represents the desired demographic groups in the latent space of interest (e.g. $\mathcal{W}^+$ latent space of StyleGANv3). Finally, we pass these latent codes to the StyleGAN's generator to obtain face images that correspond to the desired demographic group sampled from the GMMs. This process is illustrated in Algorithm 1.

Algorithm 1 Generating images of desired demographic

Input: $\mathcal{G}$, D, θ_g
Output: $\mathbf{i}_g$
$\mathbf{b} \sim \mathcal{M}(\mathbf{b}; \theta_g)$: Calculating latent according to desired demographic
$\mathbf{i}_g \leftarrow \mathcal{G}(D(\mathbf{b}))$: Generating image from the latent

4. Experiments

In this section, we describe our setup, implementation details, and various experiments that we employ to validate our results.

4.1. Validation of Synthesis

To determine if the generated images are following the desired demographic (i.e. g in $\mathbf{i}_g$ in algorithm 1), we employed an image classification task. We used the fair classifier model provided by the [28]. We used the MORPH dataset for training our autoencoders. Thus the $\mathcal{G}$ which was trained on FFHQ [31] and our autoencoder that trained on MORPH did not have any prior exposure to the images used to train the FairFace classifier.

Fig. 5 and Fig. 6 shows the confusion matrix for gender and race classification respectively. Using our method we generate *1000* image for each *male* and *female* and perform the gender classification. For race classification, we did the same with *White*, *Black* and *Latino-Hispanic*. We did not include *Asian* race demographic in this experiment as the number of samples of the MORPH dataset which we trained our autoencoder on them was to small. Also, note that the MORPH dataset only has 5 demographics for race, $\{Black, White, LatinoHispanic, Asian, Unkown\}$. The 7 classes in Fig. 6 are shown as we used the FairFace classifier. From the figures we can observe that the synthesized face images are following the group that they are sampled from.

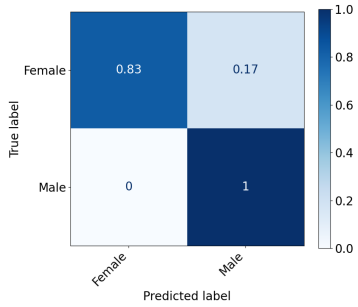


Figure 5: Confusion matrix of the gender classification task for generated images using fair classifier model

4.2. Face Recognition Experiments

Sampling demographic-specific latent $\mathbf{b}$ from a given model $\mathcal{M}(\mathbf{b}; \theta_g)$ does not necessarily guarantee that different identities will be generated. To this end, we perform FR experiment on synthesized images to verify two hypotheses:

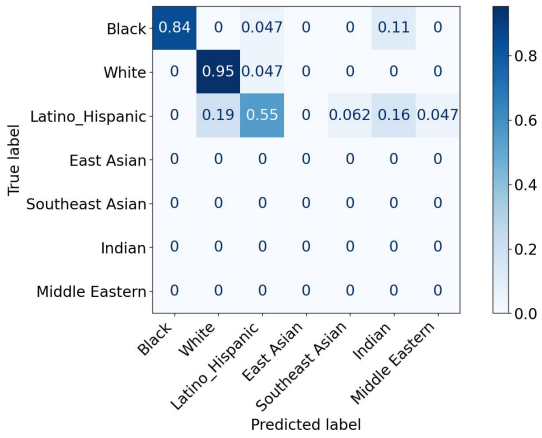


Figure 6: Confusion matrix of the race classification task for generated images using fair classifier model

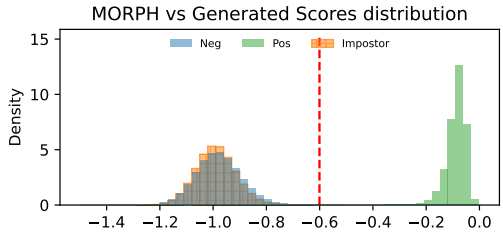
(i) the images created are part of the original data distribution of the MORPH dataset, and (ii) different identities are produced by the proposed method. The face representation is extracted using a ResNet50 network [22] trained on the WebFace4M dataset [23] using the ArcFace loss function [14]. Each pair of sample is compared using the similarity function $\mathcal{S}(u, v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2} - 1$, spanning $[-2, 0]$.

To assess that generated samples are part of the original data distribution, we compare the scores distribution of the natural image of the original MORPH dataset against the synthetically generated samples. Fig. 7a shows how the synthetic impostors (orange) compare to the real zero-effort impostors scores distribution (blue). The overlap highlights that the sampled images belong to the original data distribution and supports (i). With synthetically generated images using the proposed method, it is not possible to compare pairs of images of the same subject, as the sampling scheme does not allow to generate variability (*i.e.* pose, facial expressions, illumination) of a specific face. Therefore we can only compare the synthetic image's zero-effort impostor scores distribution to the original one to assess how different are the generated identities are.

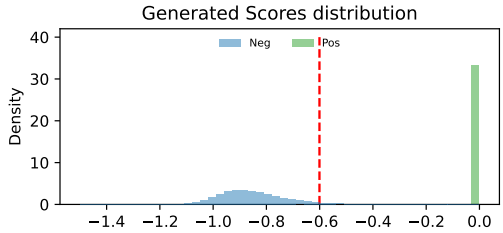
Fig. 7b shows how scores change when comparing synthetic images with themselves. The genuine score distribution is represented by a single bin because only a single synthetic image is available per identity. The zero-effort distribution (blue) moves toward the genuine score distribution (green). This shift indicates the identity difference is smaller than in the original dataset. However, the distance between the distributions remains large enough to discriminate between identities.

4.3. Demographic Preservation

Fig. 8 shows the reconstruction quality of the result of the pSp and also the reconstruction of the output of our autoencoder, more specifically, second and third columns rep-



(a) Scores of generated images against the scores of natural images of the original dataset: Genuine pairs (green), Zero-effort Impostors (blue), and Synthetic Impostors (orange).



(b) Scores of generated images: Genuine pairs (green), Zero-effort Impostors (blue)

Figure 7: Face recognition scores distributions

resent $\mathcal{G}(\mathcal{I}_{\mathcal{G}}(\mathbf{i}))$ and $\mathcal{G}(D(E(\mathcal{I}_{\mathcal{G}}(\mathbf{i}))))$ respectively. Here $\mathbf{i}$ is the original image in the dataset. Qualitatively by comparing columns in Fig. 8, we can observe that although some operations (i.e., contrastive loss) are applied to disentangle the latent space (third column), our demographic groups of interest (e.g., age, gender and race) are preserved.

4.4. Latent Space Modeling and Visualization

In this section, we show the effectiveness of our disentanglement for modeling the desired latent space.

4.4.1 t-SNE Visualization

To visually observe the complex nature for latent space of StyleGAN, we used the t-SNE plots on the test subset of MORPH dataset. In Fig. 9 we visualize the $\mathcal{W}^+$ of MORPH according to (a) gender and (b) ethnicity respectively. We can observe that gender and ethnicities according to different values are entangled and complex to model. In Fig. 10 we show effectiveness of our disentanglement method on the autoencoder's (AE) bottleneck ($\{E(\mathcal{I}_{\mathcal{G}}(\mathbf{i}_j))|\mathbf{i}_j \in \mathcal{D}_{test}\}$) and reconstruction output $\{D(E(\mathcal{I}_{\mathcal{G}}(\mathbf{i}_j)))|\mathbf{i}_j \in \mathcal{D}_{test}\}$ with (a)-(d) and without (e)-(h) our contrastive loss. We can observe that the AE's latent space with the applied contrastive loss is better disentangled according to possible demographics.

4.4.2 Likelihood Visualization

In Fig. 11, we demonstrate the modeling of different demographic groups in the latent space using likelihood plots.

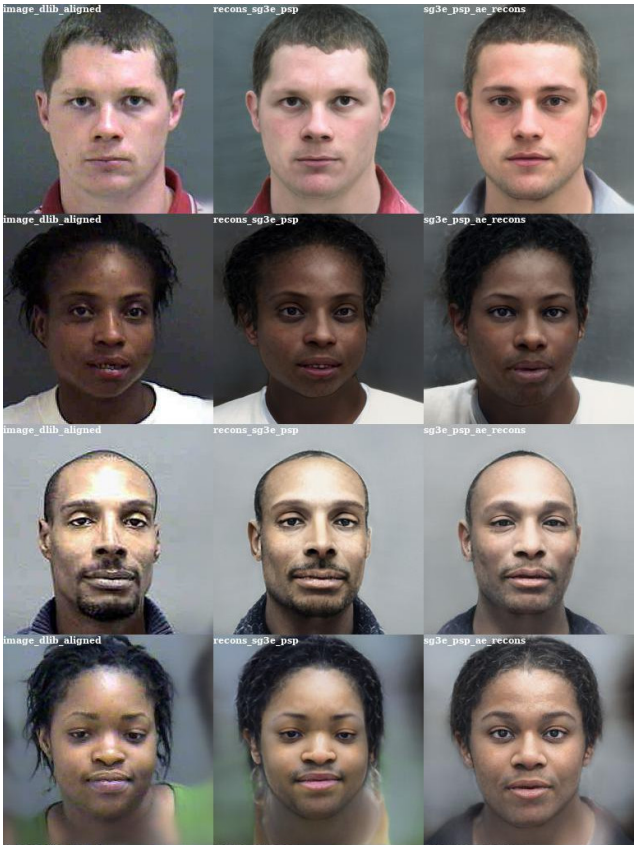
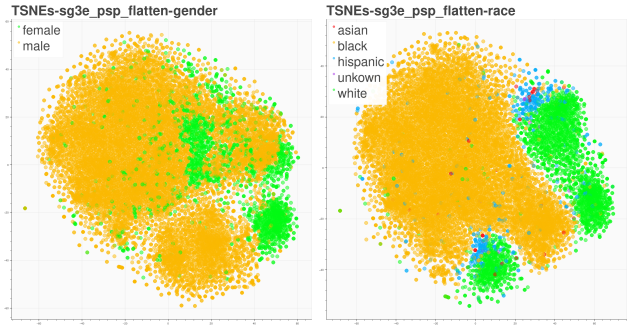


Figure 8: From left to right: the original images of MORPH dataset; reconstruction by the pSp inversion by the StyleGAN3's generator and reconstruction of the pSp inversion when passed through our disentangled autoencoder and later on passed to the StyleGAN3's generator.



(a) Original $\mathcal{W}^+$ space for *gender* demographic.

(b) Original $\mathcal{W}^+$ space for *race* demographic.

Figure 9: t-SNE plots for gender and race on the original $\mathcal{W}^+$ latent space of the StyleGANv3.

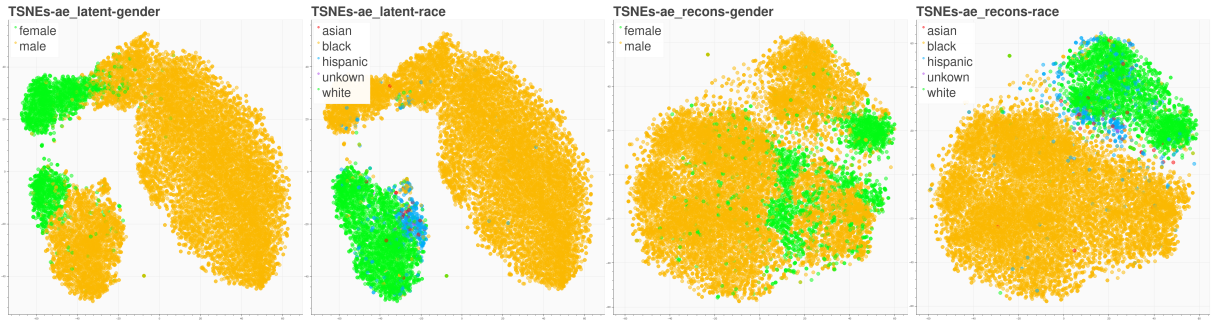
For the sake of simplicity and comparison, we limit this experiment to the gender demographic, which includes *male* and *female*. The first row represents log-likelihood plots for the original $\mathcal{W}^+$ space of StyleGAN. The first column corresponds to the LL of the model trained on train subset of the $male$ demographic in the MORPH dataset and the LL of it in comparison to the $female$ demographic of the train subset of the MORPH dataset. The second column is the same experiment except that the GMM is trained on the $female$ demographic and the LL showned in comparison with $male$ demographic. The third and fourth columns are LL plots for models trained on the previous $male$ and $female$ demographics using the train subset and the LL plots are drawn for the test subset. The second row is the same experiment settings as before beside we used the bottleneck output of our AE as modeling space. We can observe our method is effective because the overlap between two distributions ($female$ and $male$) in test cases are significantly reduced.

4.4.3 Implementation Details

We used PyTorch for our autoencoder implementation. For the trained StyleGANv3 generator and inversion based on the [49] we used the model provided by [1] paper. For the GMMs, we used scikit-learn [43]. Autoencoder was trained on a single NVIDIA RTX 3090Ti. We optimized our implementation to increase the training batch size as much as possible to minimize the effect caused by the unbalanced appearance of labels in contrastive loss. We did not change the sampling procedure to make the under-represented classes appear more frequently. We set the contribution of each demographic equally (i.e. $c_g = 1$ in Eq. 5). We experimented with different values for λ_1 and λ_2 in Eq. 6 and found that setting them to 100 and 1, respectively, worked well for a batch size of 192. We set the number of mixture components, M, to 1000. We determined this through qualitative evaluation of the reconstruction quality (e.g. using grids like in Fig. 8) as well as the contrastive loss employed in the latent space (as depicted in 11). We experimented with two versions of the autoencoder architecture: one using tensor-based encoding and decoding, and the other using a flattened version. We observed that the flattened version performed slightly better. For the encoder part, we used linear layers with dimensions of $8192-4096-2048-1024-512$, with LeakyReLU activations and an initial learning rate of 0.001. For the decoder part of the autoencoder, we employed $512 - 1024 - 2048 - 4096 - 8192$ Linear Layers with LeakyReLU activation functions for all of the layers besides the last one to preserve the range of the input-output of the autoencoder.

5. Conclusion

In this work, we present a simple yet effective method for modeling the latent-space of any StyleGAN-based generator. In contrast to previous works that are using much more complex modeling schemes we used simple modeling technique. Our method can be employed to model and later on

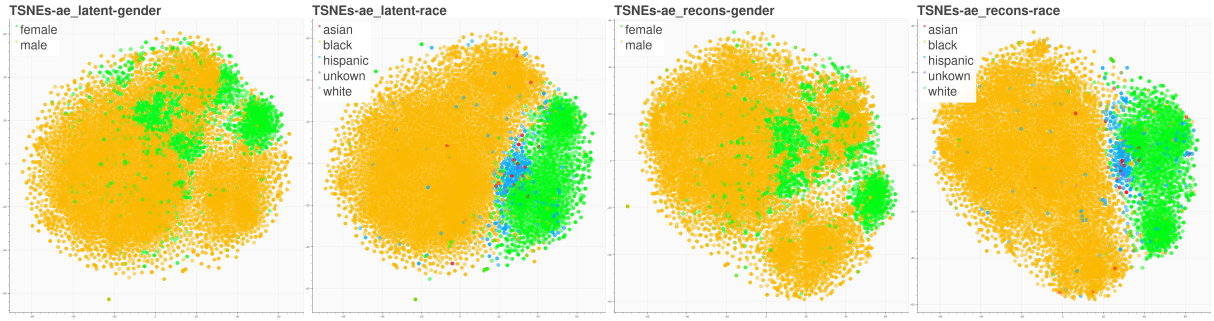


(a) AE's latent space with contrastive loss for *gender*

(b) AE's latent space with contrastive loss for *race*

(c) AE's reconstructed space with contrastive loss for *gender*

(d) AE's reconstructed space with contrastive loss for *race*



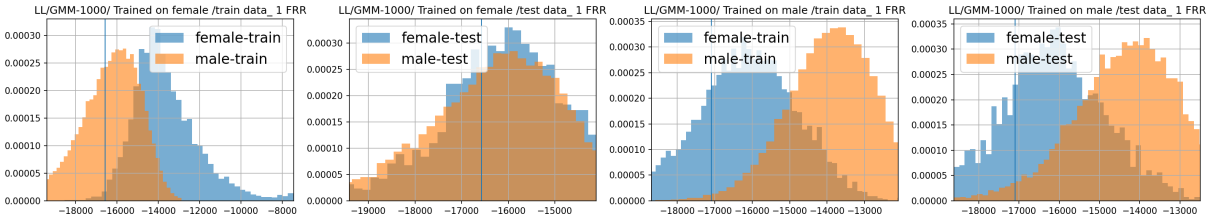
(e) AE's latent space without contrastive loss for *gender*

(f) AE's latent space without contrastive loss for *race*

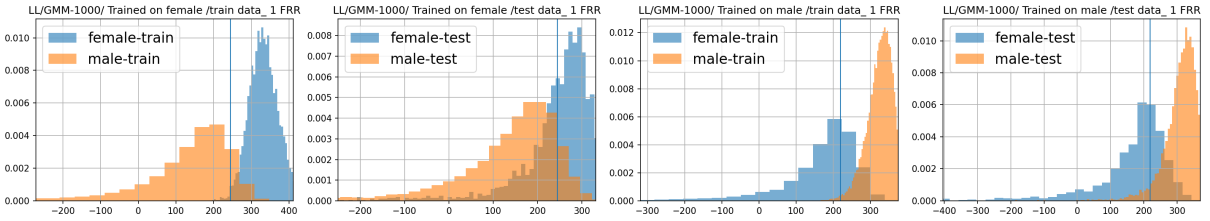
(g) AE's reconstructed space without contrastive loss for *gender*

(h) AE's reconstructed space without contrastive loss for *race*

Figure 10: t-SNE plots of various latent spaces for test part of the MORPH dataset after learning t-SNE transformation using train split of MORPH.



(a) Trained on the StyleGANs original $\mathcal{W}^+$ space



(b) GMMs trained on AEs latent space that itself trained with contrastive loss

Figure 11: Log-likelihood plots of *1000* component GMMs for various latent spaces and configurations.

generate synthetic images according to arbitrary demogrphic groups. One can categorize our proposed method as pre-processing method for addressing bias in existing models.

Acknowledgment

This research is based upon work conducted in the project SAFER and supported by the Hasler Foundation under the Responsible AI program.

References

[1] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third times the charm? image and video editing with stylegan3. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 204–220. Springer, 2023. 7

[2] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, P Winston Michalak, Shahab Asoodeh, and Flavio P Calmon. Beyond adult and compas: Fairness in multi-class prediction. *arXiv preprint arXiv:2206.07801*, 2022. 2

[3] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 3

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilovi, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019. 2

[5] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022. 3

[6] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, Jan. 2018. ISSN: 2640-3498. 1

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1

[8] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models, Jan. 2023. arXiv:2301.13188. 2

[9] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019. 2

[10] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2

[11] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on Fairness, Accountability, and Transparency*, pages 339–348, 2019. 2

[12] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020. 2

[13] Laurent Colbois, Tiago de Freitas Pereira, and Sébastien Marcel. On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021. 2

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, Long Beach, CA, USA, June 2019. IEEE. 6

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. 2

[16] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. 3

[17] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7855–7864, New Orleans, LA, USA, June 2022. IEEE. 3

[18] Yi Dong, Lei Zhen, Liao Shengcai, and Li S. Learning face representation from scratch. *ArXiv*, 2014. 1

[19] Joel Frank, Thorsten Eisenhofer, Lea Schnherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 3

[20] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 1

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[23] Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, and Tian Yang. Webface260m. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10492–10502, 2021. 1, 6

[24] EpicGames Inc. Grooming forreal-time realism:hair and fur withunreal engine. Technical report, EpicGames Inc, 2022. 3

[25] ZivaDynamics Inc. Ziva face trainer. Technical report, ZivaDynamics Inc, 2022. 3

[26] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. 2

[27] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010. 2

[28] Kimmo Karkkainen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557, Waikoloa, HI, USA, Jan. 2021. IEEE. 4, 5

[29] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 3

[30] Tero Karras, Miika Aittala, Samuli Laine, Erik Hrknen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2, 3

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3

[33] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 1

[34] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10:113–174, 04 2019. 1

[35] Blake Lemoine, Margaret Mitchell, and BrianHu Zhang. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2

[36] Jiazhi Li and Wael Abd-Almageed. Cat: Controllable attribute translation for fair facial attribute classification. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 363–381. Springer, 2023. 2

[37] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representation (ICLR)*, 2016. 2

[38] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *Advances in Neural Information Processing Systems*, pages 981–990, 2017. 2

[39] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996. 5

[40] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 4

[41] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017. 2

[42] THE EUROPEAN PARLIAMENT and THE COUNCIL OF THE EUROPEAN UNION. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, 2016. 1

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 7

[44] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[46] Behrooz Razeghi, Flavio P Calmon, Deniz Gunduz, and Slava Voloshynovskiy. Bottlenecks club: Unifying information-theoretic trade-offs among complexity, leakage, and utility. *IEEE Transactions on Information Forensics and Security (accepted)*, 2023. 2

[47] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 4

[48] K. Ricanek and T. Tesafaye. MORPH: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, Apr. 2006. 4

[49] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3, 4, 7

[50] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 42(1), aug 2022. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2

[52] Harrison Rosenberg, Brian Tang, Kassem Fawaz, and Somesh Jha. Fairness Properties of Face Recognition and Obfuscation Systems, Sept. 2022. arXiv:2108.02707. 1

[53] Mustafa Shukor, Xu Yao, Bharath Bushan Damodaran, and Pierre Hellier. Semantic unfolding of stylegan latent space. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 221–225. IEEE, 2022. 2, 4

[54] Yang Song, Hairong Qi, and Zhifei Zhang. Age progression regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 4

[55] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields, 2022. 2

[56] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2206.08361*, 2022. 2

[57] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[58] Isabel Valera, MuhammadBilal Zafar, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017. 2

[59] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018. 2

[60] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-Fidelity GAN Inversion for Image Attribute Editing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11369–11378, New Orleans, LA, USA, June 2022. IEEE. 3

[61] Qiyu Wei, Xulei Yang, Tong Sang, Huijiao Wang, Zou Xiaofeng, Cheng Zhongyao, Zhao Ziyuan, and Zeng Zeng. Latent Vector Prototypes Guided Conditional Face Synthesis. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3898–3902, Oct. 2022. ISSN: 2381-8549. 3

[62] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando D De la Torre. Generative visual prompt: Unifying distributional control of pre-trained generative models. *Advances in Neural Information Processing Systems*, 35:22422–22437, 2022. 2

[63] Weihao Xia, Yulun Zhang, Yujiu Yang, JingHao Xue, Bolei Zhou, and MingHsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2023. 3

[64] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance, 2023. 3