Demographic Bias Effects on Face Image Synthesis

Roberto Leyva
WMG, University of Warwick, UK
M.R.Leyva-Fernandez@warwick.ac.uk

Victor Sanchez
Computer Science, University of Warwick, UK
v.f.Sanchez-Silva@warwick.ac.uk

Gregory Epiphaniou
WMG, University of Warwick, UK
Gregory.Epiphaniou@warwick.ac.uk

Carsten Maple
WMG, University of Warwick, UK
CM@warwick.ac.uk

Abstract

Face image synthesis has shown remarkable progress in recent years. However, the effect that the demographics of the data used to train synthesizers has on the generation of new face images remains an open question. This paper investigates the effects of the training set demographics in the face image synthesis task. To this end, we propose a strategy that allows synthesizing face images for specific groups of people with a high visual quality. The strategy uses an unsupervised learning approach to discover groups of people in the training set based on Bayesian inference via a probabilistic mixture model. If labels are available to define the groups, our strategy can also exploit such information in lieu *of unsupervised learning. Once the groups are defined, our strategy trains a Generative Adversarial Network on each group to generate new face images with specific characteristics. Our results show remarkable performance in terms of image quality compared to several state-of-the-art baselines. More importantly, our strategy allows synthesizing face images with reduced demographic biases.*

1. Introduction

Current state-of-the-art (SOTA) synthesis methods can generate high-quality face images with fine details [16, 23]. Although some of these SOTA methods may allow for some control over the synthesis process [15], e.g., the generation of images depicting specific gestures [3, 8] or specific traits [14, 19], the influence of the training set demographics on the synthesized face images is not widely studied. This aspect is important when there is a need to synthesize face images to train models that rely on them, e.g., those that aim at distinguishing real images from fake ones. Specifically, the set of synthesized face images should be balanced in terms of its demographics to avoid introducing biases while training these models. This calls for novel methods that can synthesize face images in an unbiased manner while preserving the unique facial features of different groups of people.

In this paper, we improve our face synthesis strategy to reduce biases by accounting for the features of different groups of people . Our improved strategy relies on basic human traits to generate face images; i.e., age, gender, and race. These traits are used to define groups of people and subsequently create training subsets that can be used to train a synthesizer to generate new face images. Hence, our strategy does not focus on face trait editing via pose, geometry, or proportion, e.g., via conditional models [9, 14]. In our strategy, the groups of people can be discovered in an unsupervised manner or defined manually via ground truth labels. Our contributions are summarized as follows:

- We highlight how a synthesizer can be easily biased towards generating face images depicting a certain group of people if the training set is unbalanced.
- We synthesize high-quality face images depicting a wide range of demographics and age groups based on two widely used datasets: Flick Faces High Quality (FFHQ) and Celebrities A High Quality (CELEBA-HQ).
- We show that by appropriately defining groups of people in the training set, strong synthesis performance can be achieved even if the training samples are limited.

Our improved strategy provides the basis for fairer face image synthesis aiming at preserving the unique facial features of different groups of people. This ultimately increases trust in the face synthesis task and provides unbiased data to train other models that rely on the use of synthetic face images.

2. Related Work

Owing to the wide body of methods that are proposed to synthesize face images, we focus on those based on Generative Adversarial Network (GAN), Variational Auto Encoder (VAE), and Transformers.

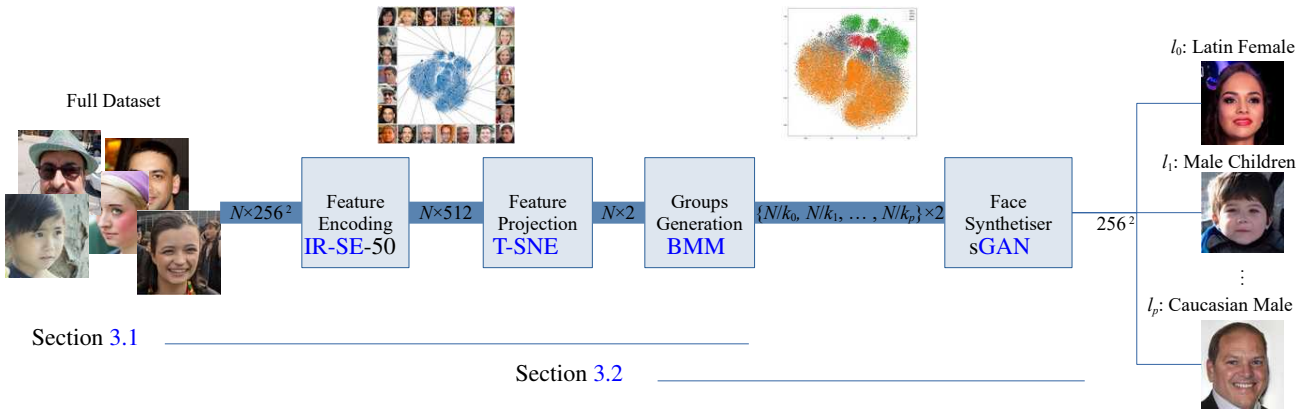


Figure 1. Proposed strategy when unsupervised learning is used to discover the groups of people. First, the training dataset comprising N images is encoded using IR-SE-50 into 512 dimensional embeddings. The dimensions of this feature space are then reduced via T-SNE. A total of K groups are discovered by using a probabilistic mixture model. A synthesizer is then trained on each discovered group.

1. Generative Adversarial Networks: much of the recent progress on face image synthesis is attributed to GANs. In [8], Gauthier *et al.* propose to condition the GAN's generation ability by creating an intermediate space that accepts noise data along with an embedding to produce an image. A two-stream GAN is proposed by Liu *et al.* [17], in which the weights of the first (last) layers of the discriminator (generator) decode (encode) high-level semantics (respectively). Such a weight-sharing constraint allows synthesizing pairs of images sharing the same level of abstraction yet having a different level of realization. Radford *et al.* [20] propose a deep CNN within a GAN framework that requires a specific formulation because CNNs are supervised models while GANs are unsupervised ones. The authors found that neither fully connected or pooling layers, commonly used in CNNs, are needed. Yin *et al.* [25] propose to traverse the latent space using a semantic definition. Their model can generate face images with specific characteristics, e.g., smiling or wearing glasses, using supervised learning. Semantics are also exploited in [5] for face image synthesis. In [12], Karras *et al.* use a coarse-to-fine GAN trained by adding layers to the generator and discriminator as the training progresses in order to generate fine details. Similarly, Struski *et al.* [26] constrain the spatial resolution to abstract local regions more accurately during the synthesis. Karras *et al.* [13] further propose to add noise and information from the latent space into the layers' blocks to improve synthesis performance.

2. Variational Autoencoders: Razavi *et al.* [21] propose an updated version of the VQ-based VAE, which relies on two deep feed-forward CNNs and requires two stages: First, a hierarchical VQ-based VAE is trained to encode images into a discrete latent space. Then, a pixel-level CNN is trained to condition the categorical distributions. Rewon *et al.* [4] show that the VAE should be as deep as the data dimensions to increase statistical dependence. Although their approach seems very computationally expensive, using one small CNN structure requires fewer parameters than other VAE methods. Vahdat *et al.* [23] propose a bidirectional encoder named Nouveau VAE, which comprises residual networks and increases expressiveness in the generated face images by partitioning the latent space.

3. Transformers: Esser *et al.* [6] propose a transformer GAN that uses the transformer's representation to quantify the vectors in the latent space generated by the VQ-based GAN [21] to learn to generate context-rich visual parts. Along the same line, Jiang *et al.* [11] propose a transformer GAN free of convolutions. Their method addresses two fundamental issues of the CNNs: their local receptive field and incapability to process long dependencies unless having several layers.

Despite advances in the face image synthesis task, only a very limited number of methods allow for some level of semantic control in the demographic generation process [5, 12, 25]. Although some methods, e.g. [14], can transfer trait attributes from the most representative groups to the synthesized face images, the synthesized face images are still biased toward these most representative groups.

3. Proposed Strategy

Figure 1 depicts the block diagram of our strategy when unsupervised learning is used to discover the groups of people present in the training dataset. After discovering several groups of people, a synthesizer is then trained on each group to generate new face images representing the corresponding group. We first explain how to define the groups by using unsupervised learning or manually based on ground truth labels, if available. Finally, we explain how new face images are generated for each defined group by using a face synthesizer.

3.1. Generation of Groups of People

The groups of people can be defined based on unsupervised learning or, if available, based on ground truth labels.

Unsupervised learning: We first generate a $d-$dimensional feature space for the training dataset by using IR-SE-50 [7, 24], which is a pre-trained model for face recognition purposes. Specifically, we use the last fully connected layer of IR-SE-50 as the feature encoder. We then project the $d-$ dimensional feature space into a $e-$dimensional space, i.e., $\mathbb{R}^d \rightarrow \mathbb{R}^e$, where $e \ll d$. After this projexction, we create a matrix $X = \{x_1, x_2, \dots x_N\}$ containing N low-dimensional samples. Finally, we train a mixture model using Bayesian inference, i.e., a Bayesian Mixture Model (BMM), with the low-dimensional data. The matrix $X = \{x_1, x_2, \dots x_N\}$ can be considered as a collection of i.i.d samples from an observable distribution. Let us define a mixture model as follows:

$$p(X|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu, \Sigma), \tag{1}$$

where $\theta = \{\pi, \mu, \Sigma\}$ is the parameter set comprising the model weights, means, and covariances, respectively, for K components. To estimate the parameters, we employ variational inference. This technique requires approximating the observed samples in X in terms of their latent variables $Z = \{z_1, z_2, \dots z_N\}$. By specifying the joint distribution $p(X, Z)$, one can estimate $p(Z|X)$ and model the evidence $p(X)$. The conditional distribution can then be written as:

$$p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n|\mu, \Lambda^{-1})^{z_{nk}}, \tag{2}$$

where $\Sigma = \Lambda^{-1}$ and z_{nk} is the n^{th} latent variable for the k^{th} component. We then estimate the variational distribution q that factorizes the latent variables and the parameters:

$$q(Z, \pi, \mu, \Lambda) = q(Z)\, q(\pi, \mu, \Lambda). \tag{3}$$

The terms on the right side of the equality in Eq. 3 can be calculated as a simplified version of the Gaussian-Wishart distribution, denoted by $\mathcal{W}$, whose scale matrix is given by W[1]. We then have:

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k \Lambda)^{-1} \mathcal{W}(\Lambda_k|W_k, \nu_k), \tag{4}$$

whose normalized estimate is given by:

$$q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{5}$$

where $\beta_k = \beta_0 + N_k$; $m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$; $\nu_k = \nu_0 + N_k$; and

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^\top, \tag{6}$$

with values given by:

$$N_k = \sum_{n=1}^{N} r_{nk}, \qquad \bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n, \tag{7}$$

$$S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^\top. \tag{8}$$

These *update* equations are analogous to the M-step equations used by the EM algorithm for the maximum likelihood solution of the mixture of Gaussians. Because a key objective of variational inference is to maximize the probability of the observed data, X, we use the component that provides the maximum posterior to define the K main groups of people:

$$k : \underset{k}{\operatorname{argmax}}\, q(Z, \pi, \mu, \Lambda). \tag{9}$$

Ground truth labels: The groups of people can be manually defined based on ground truth labels, if available. To this end, we rely on labels for *age*, *race*, and *gender*, to manually define similar groups to those discovered by unsupervised learning. As detailed in our experiments (Section 4), defining these labels manually may be challenging due to the subjectivity of these human traits.

3.2. Face Synthesizer

After the groups of people are defined, either by using unsupervised learning or by relying on ground truth labels, the final step is to train the synthesizer on each group to generate new face images for the corresponding group. We use as the backbone synthesizer the model proposed by Karras *et al.* [13] after tailoring it by reducing the input size from a 1024×1024 resolution to a 256×256 resolution. This reduction in resolution is coupled with modifications at the regularization coefficient. Because the original scale is four times the desired scale, we scale the coefficients of the original model by a factor of $4 \times 4 = 16$. Another important hyperparameter that is tailored is the number of training iterations. This number is set to 1000 using batches of 32 samples. Experimentally, we observe that the model produces acceptable results from iteration 500 upwards. It is important to mention that more recent/advanced synthesizers, e.g., diffusion models, can be used at this stage. Thus these models can be used with the proposed framework. However, we use this particular GAN-based synthesizer to conduct several experiments – as shown in the next section – in a reasonable amount of time, considering that some synthesizers require significant computing capacities [10, 22]. Hence, our strategy is synthesizer-agnostic.

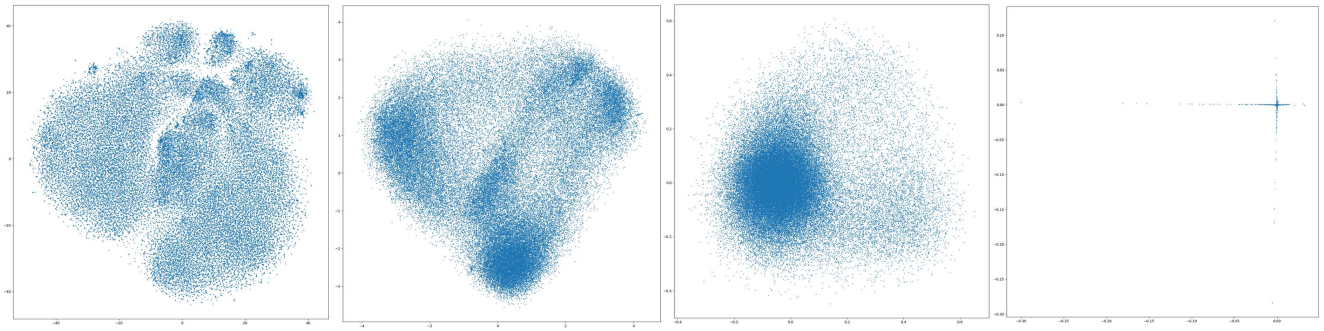


Figure 2. 2D embeddings after projecting the features computed by IR-SE-50 for the FFHQ dataset via (left to right) T-SNE, IsoMap, PCA, and LLE.

4. Experiments

We use the FFHQ [1] and CELEBA-HQ [2] datasets [12, 18] to train all synthesizers evaluated, including the one used by our strategy. In the following, we show experiments to discuss 1) the embeddings found by the feature encoder after projection to a low-dimensional space, 2) the groups discovered after training the BMM, 3) the images synthesized for the groups discovered by the BMM, 4) the images synthesized for the groups manually defined based on ground truth labels, and 5) the complexity of our strategy.

Figure 3. 2D feature space generated by IR-SE-50+T-SNE for the FFHQ dataset.

[1] github.com/NVlabs/ffhq-dataset

[2] github.com/tkarras/progressive_growing_of_gans

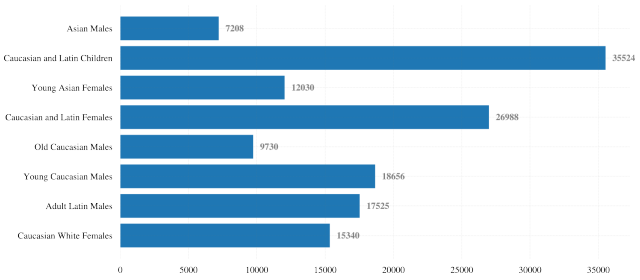


Figure 4. Dominant groups identified for the FFHQ dataset after training the BMM.

4.1. Embeddings

Figure 2 shows the resulting 2D embeddings after using different dimensionality reduction methods on the features computed by IR-SE-50. We can visually confirm that T-SNE is capable of generating the best-defined groups. Specifically, samples associated with the most dissimilar face images, e.g., the elderly and young, tend to be far from those associated with very similar faces. Conversely, IsoMaps fails to cluster samples with very similar features, creating effectively sparse groups. PCA generates only one well-defined group comprising samples with very similar features. Finally, all of the LLE embeddings produce very sparse groups and thus no visual similarity can be established.

Figure 3 shows the IR-SE-50+T-SNE embeddings in more detail. We can see that images depicting males and females are mostly situated on opposite sides of the 2D feature space. We can also observe important group formations. For instance, in the upper left region, we observe people of Asian background. While on the lower left region, we observe mostly Caucasian males. Hence, IR-SE-50+T-SNE can generate a feature space that captures the key facial characteristics of the demographic groups present in a training dataset.

(a) Caucasian females [4 left samples] and Non-Caucasian females [4 right samples].

(b) Young Males [4 left samples] Old Males [4 right samples]

Figure 5. smallSample face images synthesized by the proposed strategy using the CELEBA-HQ dataset as the training set. Each row displays samples of each of the groups discovered by the BMM.

(a) Asian males (all ages) [4 left samples], Young Latin and Caucasian females [4 right samples].

(b) Young Latin and Caucasian females [4 left samples]. Young Caucasian Males [4 right samples]

(c) Young Caucasian/Latin children [4 left samples], and Young Asian females [4 right samples].

(d) Adult Latin males [4 left samples], and Old Caucasian males [4 right samples].

Figure 6. Sample face images synthesized by the proposed strategy using the FFHQ dataset as the training set. Each row displays samples of the groups discovered by the BMM.

4.2. Groups discovered by the BMM

We quantify the number of samples assigned to each group discovered by the BMM. To this end, we visually inspect the main facial characteristics of each discovered group to assign them a label. The assigned labels and size of each discovered group are depicted in Fig. 4. Note that the *Children* group is the largest group discovered. However, we observe that children are also part of the *Asian males* group. This ambiguity is caused because IR-SE-50 produces very similar features for these two groups and hence the probabilistic model cannot discover two distinct groups for *Children* and *Asian males*. We also observe this effect in cluster boundaries, where it is difficult to correctly separate distinct groups.

Table 1. FID values of face images synthesized by the baseline and the proposed strategy using the FFHQ dataset for training.

	Proposed strategy							
Baseline[13]	Asian Males	Children	Asian Females	Latin and Caucasian Females	Old Caucasian Males	Young Caucasian Males	Latin Males	Old Caucasian Females
8.04	10.3697	6.3327	6.7907	6.6353	6.8685	6.3834	6.8685	8.2416

Results for the baseline are over all generated images as it is not trained on a per-group basis.

Table 2. FID values of face images synthesized by the baseline and the proposed strategy using the CELEBA-HQ dataset for training.

	Proposed strategy			
Baseline [13]	Non-Caucasian Females	Caucasian Females	Males	Old Males
7.79	7.8494	7.1865	8.07269	8.15650

Results for the baseline are over all generated images as it is not trained on a per-group basis.

Table 3. FID values for all face images generated by SOTA methods and the proposed strategy when using the FFHQ and CELEBA-HQ datasets for training.

Method	FFHQ	CELEBA-HQ
StyleGAN (baseline)[13]	8.04	**7.79**
VQ-VAE [21]	10.01	10.2
Ours	**7.39**	7.81

4.3. Synthesized images for groups discovered by the BMM

The face images synthesized by the proposed strategy based on the groups discovered by the BMM are compared against those synthesized by the baseline in terms of the Frechet Inception Distances (FID). This metric is useful to measure face image quality in terms of visual properties. FID values steadily increase as face images lose visual quality due to noise and distortion. Hence, low FID values are desirable [2]. Table 1 and Table 2 tabulate, respectively, the results for the face images synthesized after training the strategy on each group discovered in the FFHQ dataset and the CELEBA-HQ dataset. Notice that the proposed strategy achieves remarkable results. It is important to mention that the strategy achieves the lowest FID values for those groups with the most samples. This confirms the importance of using a balanced training set to generate new high-quality face images. From Table 1 and Table 2, one can also see that except for two groups in each case, the proposed strategy achieves better results than the baseline. Since the baseline is not trained on a per-group basis, the reported results for this baseline are for all generated images.

We visually inspect the synthesis results to corroborate the performance per group of people. Figures 6 and 5 show sample synthetic face images generated for each of the discovered groups. The dominant groups identified in the FFHQ dataset are *Asian males* (all ages), *Young Latin and Caucasian females*, *Old Caucasian females*, *Young Caucasian males*, *Young Caucasian and Latin children*, *Young Asian females*, *Adult Latin males*, and *Old Caucasian males*. The dominant groups identified in the CELEBA-HQ dataset are *Caucasian females*, *Non-Caucasian females*, *Old males*, and *Males*. One can see that the proposed strategy can accurately generate face images for each group preserving the group's main facial features.

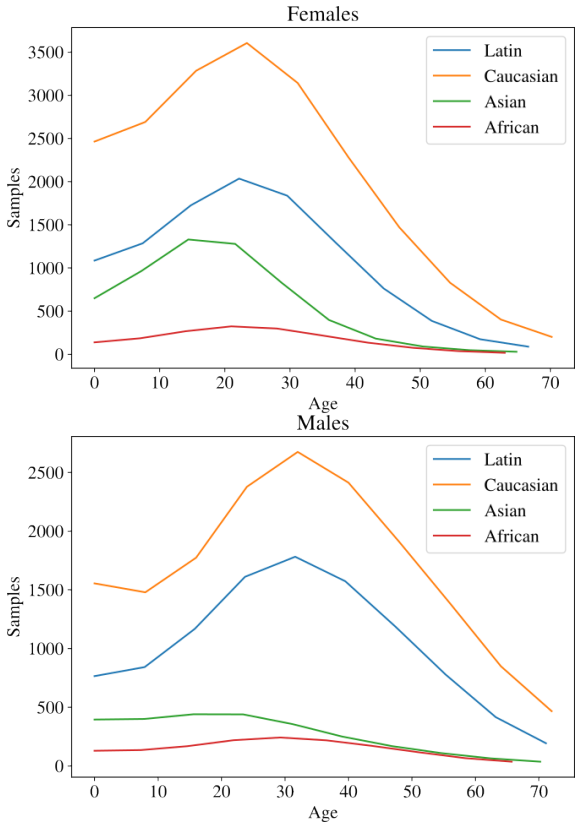


Figure 7. Label distribution for the FFHQ dataset.

Table 3 tabulates FID values for all face images generated by two SOTA methods and our strategy using the FFHQ and CELEBA-HQ datasets for training. We can see that our strategy outperforms these SOTA methods when using the FFHQ dataset for training and attains very competitive performance when using the CELEBA-HQ dataset for training.

Table 4. FID values for all face images generated by the baseline and those generated by the proposed strategy on a per-group basis using the FFHQ dataset for training.

	Proposed strategy			
Baseline [13]	Caucasian Males (Ages 25:99)	Caucasian Females (Ages 25:99)	Latin Males (Ages 16:60)	Latin Females (Ages 16:60)
8.04	5.9367	5.4280	6.9149	6.2536

Results for the baseline are overall generated images as it is not trained on a per-group basis.

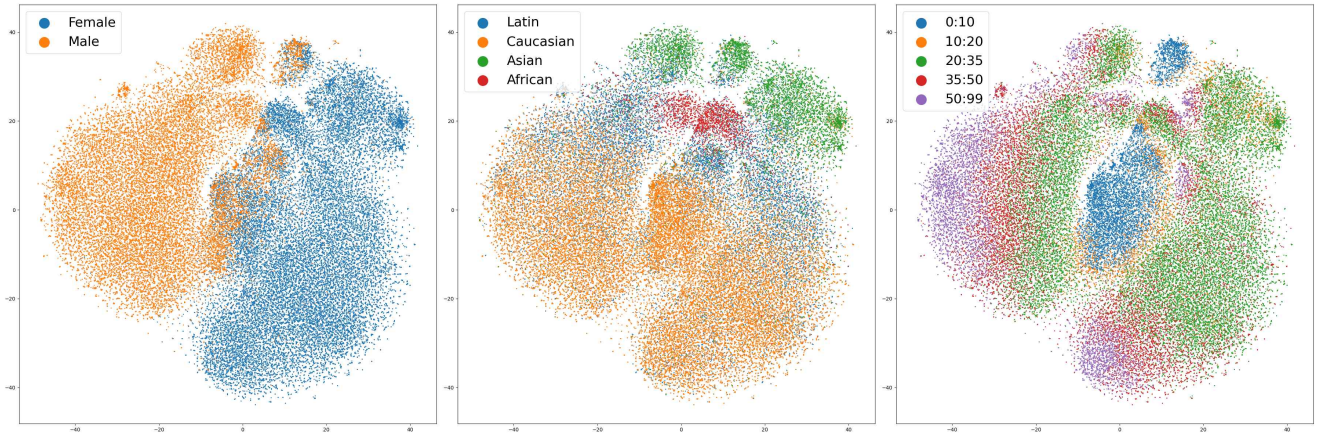
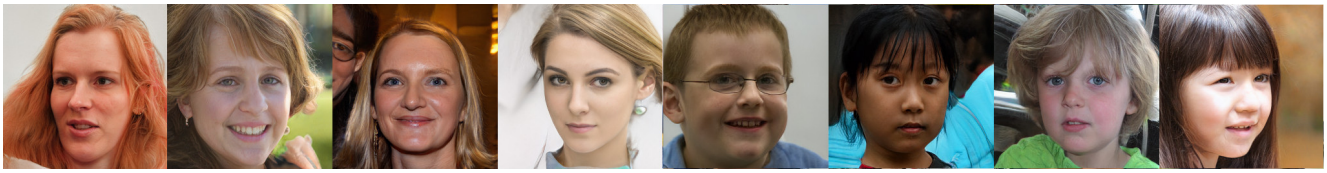


Figure 8. 2D feature space generated by IR-SE-50+T-SNE for the labeled FFHQ dataset (best viewed in color). (Left) Samples by gender. (Centre) Samples by race. (Right) Samples by age group.

(a) Latin females (all ages) [4 left samples]. Asian females [4 right samples]



(b) Caucasian females (ages 15 – 50) [4 left samples], and Children (ages 0 – 10) [4 right samples].

Figure 9. Sample synthesized images by the proposed strategy using the FFHQ dataset as the training set. Each row displays samples of groups manually defined based on ground truth labels.

4.4. Synthesized images for manually defined groups

We label the 70,000 samples of the FFHQ dataset. The labels correspond to the *age*, *race*, and *gender* of the depicted faces. Figure 7 shows the label distribution, from which we can see that the FFHQ dataset comprises mainly face images of females between the ages of 16 and 40. It is worth noting that the *race* label is hard to manually define due to the inherent difficulty of determining someone's race purely based on their face image. For instance, Latin, African, and Middle Eastern people share very similar facial characteristics, making it very difficult to distinguish them. The labeling process for the *race* label resembles the process of determining one person's nationality. We then opt for creating groups as large as possible whose members share the most distinctive traits, specifically, skin complexion, hair color, eyes, and overall face shape. We then define four main *race* labels: *Caucasian*, *Latin*, *African*, and *Asian*.

Figure 8 shows the 2D feature space computed by T-SNE for the labeled FFHQ dataset based on the feature vectors given by IR-SE-50+. One can see that males and females (Fig. 8-left) can be easily separated into two clusters that overlap in the mid-section of the 2D space. Interest-

ingly, this overlapping section corresponds to children, thus it is understandable why people may confuse the gender in such cases. One can also see that the *Asian* and *Caucasian* groups (Fig. 8-center) are effectively located on opposite sides of the 2D space. We observe that overlapping areas correspond to face images whose race cannot be easily inferred. Regarding the *age* label (Fig. 8-right), we observe that the embeddings also provide useful information: the young and elderly are clearly separated in the 2D space.

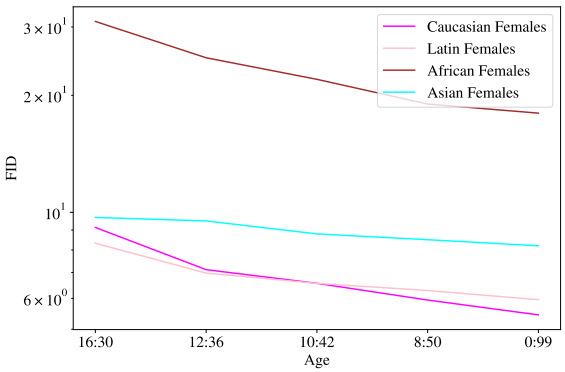


Figure 10. FID values for the face images synthesized by the proposed strategy for each group defined based on the *race* label as a function of the age range.

Based on the manually-defined groups, we train the synthesizer on each group and evaluate the synthesized face images generated for each group. Figure 9 shows sample face images synthesized for the groups with the most samples. i.e., the *Female* and *Children* groups. We can observe that the synthesized images accurately resemble the characteristics of the corresponding group of people when the demographics are taken into account for training. The quality of the synthesized images is particularly high in cases where the number of training samples is large. This emphasizes the need to have enough samples from all groups to generate new images in an unbiased manner. For example, we observe that the *African males* and the *Asian males* groups have significantly fewer training samples than the *Caucasian males* group. Thus the baseline, which is trained on all training images, rarely generates new faces for these two groups.

Figure 10 shows FID values of the face images synthesized by our strategy when trained on the four groups defined based on the *race* label for several age ranges. As expected, as the age range is widened, the FID values decrease and thus the synthesizer generates less distorted face images because more training samples are available. We observe that for some groups, the synthesizer requires fewer training samples to generate high-quality face images. Specifically, for the *Caucasian females* and *Latin females* groups, nearly the same FID value is achieved for the 10:42 age range even though the former group has nearly 50% more training samples than the latter. Nevertheless, we observe that, in general, more training samples improve the synthesis process. For example, the face images generated for the *African females* group, which is the group with the smallest number of training samples, tend to have the lowest quality.

The last experiment trains the proposed strategy after varying the number of training samples for specific groups. Table 4 tabulates the results for two well-populated (*Caucasian Males* and *Caucasian Females* older than 25 years) and two fairly-populated (*Latin Males* and *Latin Females* older than 25 years) groups. We observe that the strategy outperforms the baseline for these four groups, particularly for the first two groups. We can then argue that 1) defining the groups of people appropriately helps to improve the synthesis performance. And 2) strong performance can be achieved even if there are not many training samples for a group but the group is defined based on relevant and adequate labels.

5. Conclusions

This paper presented a strategy to generate face images for different groups of people by accounting for the similarities of the training samples, in terms of basic human traits. The paper also showed the effects of the under-representation of groups of people in the training set on the face synthesis task. Based on extensive experiments, we showed that some groups of people may require more training samples than others for a synthesizer to achieve the same quality in the synthesized face images. These results reveal that reducing biases in the face image synthesis process is more elaborate than just balancing the training set. Finally, our results showed that the proposed strategy can attain strong performance in terms of image quality compared to SOTA methods. As part of our future work, we will continue improving our strategy to synthesize high-quality face images for highly under-represented groups in the most common training datasets, e.g., African males, by inspecting closely the latent space of the available training samples. We will also investigate more closely the relation between number of training samples available for each group of people and the quality of the corresponding synthesized face images.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 3

[2] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 6

[3] Behzad Bozorgtabar, Mohammad Saeed Rad, Hazım Kemal Ekenel, and Jean-Philippe Thiran. Learn to synthesize and synthesize to learn. *Computer Vision and Image Understanding*, 185:1–11, 2019. 1

[4] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020. 2

[5] Chris Donahue, Zachary C Lipton, Akshay Balsubramani, and Julian McAuley. Semantically decomposing the latent spaces of generative adversarial networks. *arXiv preprint arXiv:1705.07904*, 2017. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[7] Lu Feihong, Chen Hang, Li Kang, Deng Qiliang, Zhao jian, Zhang Kaipeng, and Han Hong*. Toward high-quality face-mask occluded restoration. *T-OMM*, 2022. 3

[8] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014. 1, 2

[9] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3436–3445, 2019. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[11] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2, 4

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228, 2021. 2, 3, 6, 7

[14] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023. 1, 2

[15] Zhenzhong Kuang, Zhiqiang Guo, Jinglong Fang, Jun Yu, Noboru Babaguchi, and Jianping Fan. Unnoticeable synthetic face replacement for image privacy protection. *Neurocomputing*, 457:322–333, 2021. 1

[16] Feng Liu, Hanyang Wang, Jiahao Zhang, Ziwang Fu, Aimin Zhou, Jiayin Qi, and Zhibin Li. Evogan: An evolutionary computation assisted gan. *Neurocomputing*, 469:81–90, 2022. 1

[17] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 0–0. Curran Associates, Inc., 2016. 2

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 4

[19] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2023. 1

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[21] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2, 6

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[23] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 1, 2

[24] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face. evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021. 3

[25] Weidong Yin, Yanwei Fu, Leonid Sigal, and Xiangyang Xue. Semi-latent gan: Learning to generate and modify facial images from attributes. *arXiv preprint arXiv:1704.02166*, 2017. 2

[26] Łukasz Struski, Szymon Knop, Przemysław Spurek, Wiktor Daniec, and Jacek Tabor. Locogan — locally convolutional gan. *Computer Vision and Image Understanding*, 221: 103462, 2022. 2