Zero-Shot Demographically Unbiased Image Generation From an Existing Biased StyleGAN

Anubhav Jain, Rishit Dholakia, Nasir Memon, *Fellow, IEEE*, and Julian Togelius

***Abstract*—Face recognition systems have made significant strides thanks to data-heavy deep learning models, but these models rely on large privacy-sensitive datasets. Recent work in facial analysis and recognition have thus started making use of synthetic datasets generated from GANs and diffusion based generative models. These models, however, lack fairness in terms of demographic representation and can introduce the same biases in the trained downstream tasks. This can have serious societal and security implications. To address this issue, we propose a methodology that generates unbiased data from a biased generative model using an evolutionary algorithm. We show results for StyleGAN2 model trained on the Flicker Faces High Quality dataset to generate data for singular and combinations of demographic attributes such as *Black and Woman*. We generate a large racially balanced dataset of 13.5 million images, and show that it boosts the performance of facial recognition and analysis systems whilst reducing their biases. We have made our code-base (https://github.com/anubhav1997/youneednodataset) public to allow researchers to reproduce our work.**

***Index Terms*—Bias mitigation, face recognition, synthetic datasets.**

I. Introduction

FACE recognition systems that were once based on handcrafted features have now achieved human-level performance with the assistance of deep learning models. However, this transition has resulted in the accumulation of large privacy-sensitive datasets that are costly to collect and pose several issues. One major issue is that these large datasets often lack ethnic and demographic diversity, which causes deep facial recognition models to suffer from similar biases. Ensuring the collection of highly diverse image datasets is not only difficult but also expensive. Another issue is that many countries have recognized biometric data privacy as a fundamental right and have regulated its collection and usage by law [1], [2], [3]. This makes it challenging to collect data from a large number of users and raises privacy concerns. Companies like Facebook [4], Google [5], and Shutterfly [6] have faced scrutiny for their usage of facial images of users under the BIPA law.

This work presents an approach to address the issues of bias and data privacy in facial recognition models by leveraging advancements in image generation. Generative models offer a cost and time-effective alternative to manual data collection and annotation. They provide better control over environmental conditions, lighting, occlusions, camera angles, and backgrounds, enabling the training of more robust and adaptable models. Additionally, synthetic datasets do not contain any personally identifiable information, thus providing crucial privacy protection. There are also various applications related to identity anonymization where controlled facial and protected attributes are required.

However, there is a major challenge in using existing generative models as they are highly biased. For instance, when randomly sampling 10,000 images from a StyleGAN2 [7] model, only 26 corresponded to Indians, 171 to Africans, while over 6500 were Caucasians, as depicted in Figure 2. Thus, generating large balanced datasets through simple rejection sampling is not only inefficient but also implausible for underrepresented groups.

Previous methods that aimed to generate data for specific protected attributes, such as race, have either trained a generative model from scratch or fine-tuned an existing model. However, both of these approaches either require the collection of large amounts of balanced real data or depend on its availability. For instance, a recent study [8] collected over 5 million images of Africans and 3 million images of Asians from various YouTube sources. Similarly, [9] used real datasets such as FairFace, UTKFace, and MORPH. Reference [10] instead generated 256,000 synthetic images just to find latent directions pertaining to different demographic groups. In contrast, we propose a simple yet effective search-based algorithm that exploits the disentangled nature of the StyleGAN latent space. This approach can generate a large number of demographically balanced unique synthetic identities in a zero-shot manner, i.e., without any training or using real datasets.

Moreover, as compared to optimizing over a single protected attributes as in [8], the ability to generate combination of protected attributes allows jointly optimizing for multiple fairness objectives. Though, it is important to note this is limited by the existence of such variations in the existing StyleGAN2 model. We show examples of the output of the proposed algorithm in Figure 1.

Manuscript received 15 November 2023; revised 23 May 2024; accepted 12 June 2024. Date of publication 18 June 2024; date of current version 18 November 2024. This work was supported by NSF under Grant 1956200. This article was recommended for publication by Associate Editor G. A. Rocha upon evaluation of the reviewers' comments. *(Corresponding author: Anubhav Jain.)*

Anubhav Jain and Julian Togelius are with the Tandon School of Engineering, New York University, New York, NY 11201 USA (e-mail: aj3281@nyu.edu; jt125@nyu.edu).

Rishit Dholakia is with the Courant Institute of Mathematical Sciences, New York University, New York, NY 11201 USA (e-mail: rnd7446@nyu.edu).

Nasir Memon is with the Computer Science, Data Science, and Engineering Department, New York University, Shanghai 200126, China (e-mail: memon@nyu.edu).

Digital Object Identifier 10.1109/TBIOM.2024.3416403

Fig. 1. Examples of images generated using our approach for combination of protected attributes.

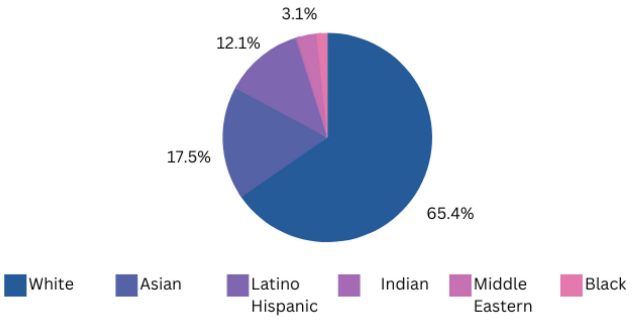


Fig. 2. Pie-chart depicting the racial distribution in the generated output of the StyleGAN2 model across 10000 randomly drawn samples.

To show how this approach could be used to boost facial recognition and analysis networks while mitigating biases, we generate a racially balanced synthetic dataset. The dataset consists of 50,000 synthetic identities each for six different racial groups, including Indian, White, African, Black, Asian, Middle Eastern, and Latino Hispanic, resulting in a total of 13.5 million images with 45 images per person. We demonstrate that pretraining on this data can significantly improve the performance of facial recognition systems. We evaluate three different systems, namely ArcFace [11], AdaFace [12], and ElasticFace [13], and show that our approach outperforms models trained using widely accepted unbalanced datasets such as VGGFace2 [14]. Our approach shows improvements even on the BUPT balanced-face dataset [15] which already contains an equal number of identities per race. Similarly, we show results on ethnicity and gender classification using the same generated dataset on the FairFace [16] and the UTKFace datasets [17]. We show that the approach helps in boosting performance while reducing biases.

To summarize, we make the following contributions in this paper:

- We propose a simple evolutionary search-based approach to generate a large balanced set of images using an existing biased generative model. Our approach doesn't require any training dataset, synthetic or real. Additionally, it doesn't require training or fine-tuning of the generative model.
- We contribute a dataset of over 50,000 distinct synthetic identities for six different racial groups resulting in a total of 13.5 million images with 45 images per person.
- We show that pre-training on the generated dataset improves the performance of facial recognition and analysis models while reducing the bias present in the models.
- We contribute to the ongoing discussion of addressing bias in facial recognition by providing a practical and scalable solution that can be implemented without requiring the collection of large amounts of real data.

This paper was invited to be submitted as an extension of our previous IJCB 2023 paper [18], and compared to that paper this manuscript contains the following additional material:

- We extend the work to show we use an modified fitness function of the evolutionary algorithm to generate face images with a combination of protected attributes.
- We further expanded on the explanation and analysis of the proposed approach.
- We show that the generated racially balanced dataset can be used to boost facial analysis algorithms while reducing biases. We show results on the FairFace and UTKFace datasets for ethnicity and gender classification tasks.
- We extended the training framework of face recognition systems on the generated dataset to show results when using only 4 racial groups, as is the case with the Racial Faces in the Wild (RFW) dataset.

II. Related Work

This section provides an overview of the most relevant work done on bias mitigation in facial recognition models, including the use of synthetic data for training these models, which is particularly pertinent to the proposed approach.

A. Bias Mitigation in Face Recognition

Past research has extensively shown that widely accepted deep learning-based facial recognition algorithms exhibit bias towards a particular ethnicity [19], [20], [21], [22]. Most of the research done on bias mitigation has been directed towards mitigating the bias for particular demographic subgroups rather than arriving at generalizable solutions across demographic groups [19].

Zhang et al. [23] proposed an adversarial learning approach to reduce bias in facial recognition systems. Yucer et al. [24] proposed an approach to alter the ethnicity of a person through an adversarial training procedure applied to a CycleGAN model. Gong et al. [25] used adaptive convolutional kernels

and attention mechanisms based on the demographic subgroup to mitigate demographic biases. Wang and Deng [15] propose a reinforcement learning-based race balance network. They also introduce the BUPT-GlobalFace and BUPT-BalancedFace datasets containing datasets with racial distribution on global and balanced distributions respectively. Researchers have also proposed approaches to remove protected attributes in data representations by adversarial training models [23], [26], [27].

Other researchers who studied biases in facial recognition models have pointed out correlations between protected attributes and other features which is often the reason for biases in face recognition models. Researchers have proposed approaches to disentangle the protected attribute from other features [28] as well as suppressing the protected attribute [29] to have fairer face recognition models. In our case, given the control that generative models provide, we can ensure that some attributes such as pose, illumination, and expression do not correlate with the protected attribute.

B. Synthetic Data Generation

Recently, researchers have shown interest in the use of synthetic data for face recognition systems due to their privacy-preserving properties. Most researchers have trained new generative models on different types of real datasets [8], [30], [31], [32]. Boutros et al. [31] proposed an approach to train facial recognition using synthetic images. They trained a generative model conditioned on the user identity to create a synthetic dataset. Sevastopolsky et al. [8] also proposed an approach to train a generative model on unlabelled data collected from YouTube. They use this model to train an encoder model which is then finetuned for face recognition. However, all of these methods require a large number of either labeled or unlabelled images for training the generative models for their task. In this paper, we eliminate this step by making use of an existing generative model even though it is biased.

Ramaswamy et al. [33] proposed a method to de-correlate target labels (e.g., glasses, hats) with protected attributes (e.g., race, gender) to remove biases in facial attribute classification models. Other researchers have also proposed approaches to alter the facial attributes in images using generative models [34], [35], [36], [37], [38]. Researchers have also found ways to disentangle the identity of a person from other facial attributes of the image [39], [40].

A more related line of research to our work has been on using pre-trained generative models to edit images by traversing the latent space. Colbois et al. [41] demonstrated that specific latent directions exist in the StyleGAN2 latent space that can modify the pose, illumination, and expressions of synthetic identity. Similarly, [10] used synthetic images labeled by an auxiliary classifier as feedback to find latent directions corresponding to different facial and protected attributes. Reference [9] instead proposed using Gaussian mixtures on a disentangled lower dimensional space instead of directly using the StyleGAN latent space. Interestingly, they claimed the StyleGAN3 latent space was not disentangled for protected attributes. This could be because, as [42] claimed, the StyleGAN3 latent space is more entangled than the StyleGAN2 counterpart. We discuss this further in Appendix 2. Jain et al. [18] showed that the subspaces exist in the original latent space of the StyleGAN2 model and can be explored by using an evolutionary algorithm.

(a)

(b)

Fig. 3. We produce intra-class variations in expression (first row), poses (second row) and illumination (third row) using latent directions in the StyleGAN latent space. As we show, these variations work well across demographics while preserving the identity.

III. Unbiased Data Generation

In this section, we present our approach to generate demographic-specific data. We start by defining the problem in Section III-A followed by the proposed evolutionary algorithm in Section III-B.

A. Problem Statement

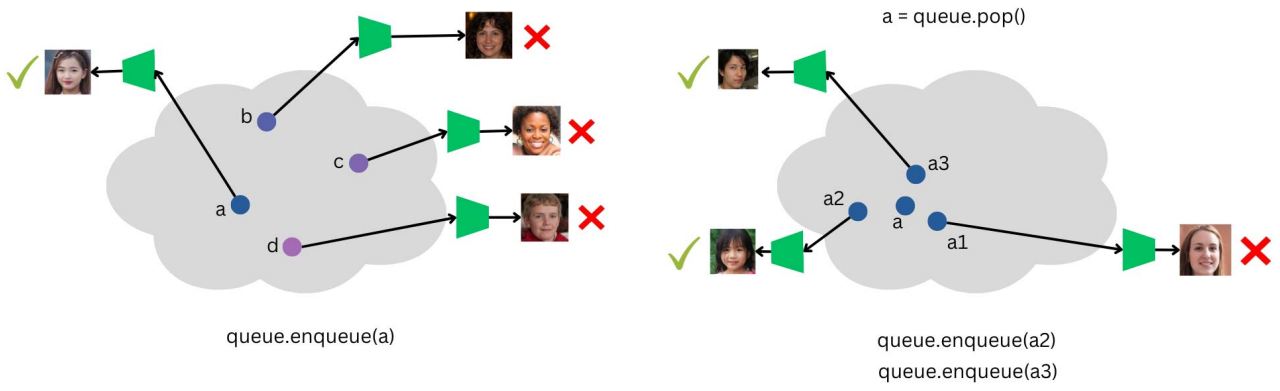
We are given a generative model that exhibits biases wrt to certain demographic groups and we wish to generate unbiased data from this model. Let us assume a generative model $\mathcal{G}$ parameterized by θ and learned distribution $p_\theta : \mathcal{X}, \mathcal{D} \rightarrow \mathbb{R}$ over a set of demographic groups $\{d_1, d_2, \ldots, d_n\} \in D$ and samples $x \in \mathcal{X}$. $p_\theta(d_i)$ is the marginal distribution over the demographic class d_i for the joint distribution $p_\theta(x, d)$. Thus,

$$p_\theta(d_i) \neq p_\theta\big(d_j\big) \quad \forall d_i, d_j \in \mathcal{D}; i \neq j. \tag{1}$$

$$\mathbb{E}_{x\sim p_\theta}[x \in d_i] \neq \mathbb{E}_{x\sim p_\theta}\big[x \in d_j\big] \quad \forall d_i, d_j \in \mathcal{D}; i \neq j. \tag{2}$$

This implies that in expectation when randomly sampling data from the generative model we do not get an equally representative set across the group of demographics. We wish to mitigate this disparity such that, for a generated dataset with data distribution $p_{gen} : \mathcal{X}, \mathcal{D} \rightarrow \mathbb{R}$,

$$\mathbb{E}_{x\sim p_{gen}}[x \in d_i] = 1/|D| \quad \forall d_i \in D. \tag{3}$$



(a) Random Rejection Sampling (b) Iterative search in the subspace

Fig. 4. Our proposed approach consists of two parts, (a) randomly sampling latent vectors till demographic criteria are met, followed by (b) iteratively sampling in the subspace of the found vector.

(a) Seed (b) Iteration 1 (c) Iteration 2 (d) Iteration 3 (e) Iteration 4 (f) Iteration 5

Fig. 5. Images showing the generation of synthetic data using our approach targeting the 'Latino Hispanic' ethnic group and the 'Woman' gender group while mutating in the z-subspace with the number of mutations set to 3.

In this work, we use a pre-trained StyleGAN2 [7] model that was trained on the Flickr-Faces High-Quality [43] dataset to generate 1024x1024 closeup facial images. The StyleGAN2 model uses two latent spaces, the space $\mathcal{Z} \in \mathbb{R}^{512}$ sampled from a Gaussian random vector. This is mapped to a larger latent space $\mathcal{W}+ \in \mathbb{R}^{18\times512}$ using a small neural network-based mapping function.

B. Generating Demographic-Specific Identities

While there exists representation biases in the StyleGAN model, there is still significant variability which can allow the generation of unique synthetic identities that are balanced in terms of the demographic groups. Reference [44] showed that StyleGAN models have a capacity of approximately 1.43×10^6 unique identities. Additionally, previous research has shown the existence of latent directions or subspaces in the StyleGAN2 latent space pertaining to different facial or protected attributes [10], [41]. In this work, we make use of these subspaces to propose using a controllable latent space search algorithm similar to a breadth-first search in the latent space of a StyleGAN2 model. Figure 4 presents a high-level overview of the proposed approach which consists of two parts. The first step is finding a latent vector in the subspace pertaining to the target demographic group. This is done by random rejection sampling using auxiliary race, gender and age classifiers [45] for checking whether the generated image belongs to the target groups. Using these demographic classifiers we define a fitness function as follows,

$$f(\mathbf{v}) = \begin{cases} 1 \; if \;\; \mathcal{C}_r(\mathcal{I}), \mathcal{C}_g(\mathcal{I}), \mathcal{C}_a(\mathcal{I}) == T_r, T_g, T_a \\ 0 \; otherwise \end{cases} \tag{4}$$

where T_r, T_g, T_a are the target demographic groups corresponding to gender, age and ethnicity, $\mathcal{G}$ is the generative model, and $\mathcal{C}_r, \mathcal{C}_g, \mathcal{C}_a$ are the race, gender and age classifiers respectively. We use the latent vector found through random sampling as the starting point, referred to as $\mathbf{v}_s$ in Algorithm 1. Similar to the breadth-first search we maintain a queue for the traversal. The starting point is added to a queue and we use this to begin the search.

Iteratively, we dequeue a latent vector from the queue, referred to as $\mathbf{v}_c$. If $f(\mathbf{v}_c) = 1$ then, we sample neighboring points by mutating the current latent vector $\mathbf{v}_c$ with a random variable. We use a uniform random variable instead of a multivariate Gaussian random variable, which is typically used in most search algorithms, as it provides better control in the GAN latent space and allows us to generate reasonable facial images by staying within appropriate boundaries. Let $\mathbf{v}_i$ be the set of vectors mutated from the vector $\mathbf{v}_c$ at the i-th iteration.

$$\mathbf{v}_i = \{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{in}\} \tag{5}$$

where n is the number of mutations of the current vector $\mathbf{v}_c$. We append the entire set $\mathbf{v}_i$ into the queue. We continue the search process till the queue is not empty and other search controls such as maximum iterations have not been exceeded. We show the iterative generation of synthetic identities in Figure 5 for the Asian demographic subgroup with the number of mutations set to 1.

Our approach works well on both the $\mathcal{Z}$ and $\mathcal{W}+$ latent vector space, even though previous research has suggested limited disentanglement of the $\mathcal{Z}$ space. The trade-off here is similar to the use of a truncation-psi parameter, between the diversity and quality of the images. The $\mathcal{Z}$ latent space

Algorithm 1 Latent Space Exploration for Generating Synthetic Identities for Each Demographic Subgroup

Input: A generative model $\mathcal{G}$; an auxiliary race, gender and age classifier $\mathcal{C}_r$, $\mathcal{C}_g$, $\mathcal{C}_a$; target demographic group T_r, T_g, T_a; starting latent vector found using random sampling $\mathbf{v_s}$ s.t. $\mathcal{C}(\mathcal{G}(\mathbf{v_s})) = t$; number of mutations n; max range of random mutation δ; maximum number of iteration for a particular starting vector *max_iter*

bfOutput: A list of latent space vectors $\mathbf{w}$.

queue $\leftarrow$ []
out $\leftarrow$ []
iter $\leftarrow 0$
queue.enqueue($\mathbf{v_s}$)
while *len*(*queue*) $\neq 0$ and *iter* $\leq$ *max_iter* **do**
  $\mathbf{v_c} =$ *queue.dequeue*()
  $\mathcal{I} = \mathcal{G}(\mathbf{v_c})$
  if face not detected in $\mathcal{I}$ **then**
    continue
  end if
  if $\mathcal{C}_r(\mathcal{I}) == T_r$ and $\mathcal{C}_g(\mathcal{I}) == T_g$ and $\mathcal{C}_a(\mathcal{I}) == T_a$ **then**
    out.append($\mathbf{v_c}$)
    iter $\leftarrow$ *iter* $+ 1$
    for j $= 0$ **to** n **do**
      $\mathbf{v_j} \leftarrow \mathbf{v_c} +$ *random*(*range* $= [-\delta, \delta]$)ju
      if *dist*($\mathbf{v_j}$, $\mathbf{v_s}$) $>$ *dist*($\mathbf{v_c}$, $\mathbf{v_s}$) **then**
        queue.enqueue($\mathbf{v_j}$)
      end if
    end for
  end if
end while
return List of latent vectors corresponding to the target ethnicity - *out*.

ensures better quality, however with limited diversity. In contrast, the $\mathcal{W}+$ latent space ensures higher diversity but may compromise the quality of the image. Thus, we use a Google mediapipe [46] face detection model when searching in the $\mathcal{W}+$ latent space to ensure we stay within bounds. This is however not required when evolving the $\mathcal{Z}$ latent space which follows a smooth Gaussian distribution. Also, in the case of mutating in the $\mathcal{W}+$ latent space, we observed that after a large number of iterations (> 500), when a number of possible directions had been exhausted, the synthetic identities started looking similar. This is because the search algorithm is forced to take directions where the identity doesn't change but only variants of the same identity are produced. We see that limiting the number of iterations from a particular seed value can efficiently take care of this problem.

The proposed approach runs independently for different ethnicities as shown in Algorithm 1 and can thus parallelly generate data for different ethnicities. This allows us to specifically control how many samples are required for each demographic subgroup and we can appropriately terminate the search operation once this criterion is satisfied.

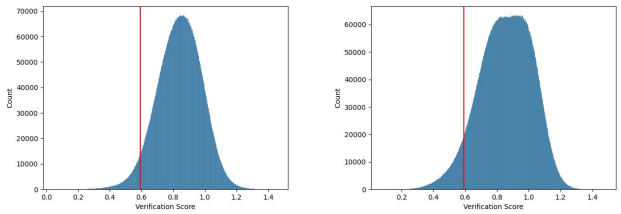
For generating multiple images for each identity, we use the approach proposed by Colbois et al. [41] using latent directions for generating expression, pose, and illumination variations. We have shown examples of intra class variations in Figure 3. We have, however, excluded extreme pose variations due to the limitations of the approach on StyleGAN2 as shown by the original authors. This allows us to specifically curate a diverse yet controlled set of facial expressions, poses, and illumination while maintaining consistency across identities and ethnicities.

To show advantages of the approach in face recognition and analysis we use the proposed approach to generate a large racially balanced dataset. We have broadly classified images into 6 ethnic groups using an auxiliary ethnicity classifier [45] - Caucasian, African, Indian, Asian, Middle Eastern, and Latino Hispanic. In comparison to previous studies that have generally used only 4 racial groups, we believe this is more inclusive even though the test face recognition datasets only contain labels for 4 groups - Indian, African, Caucasian, and Asian. We generate two versions each for the $\mathcal{Z}$ and $\mathcal{W}+$ spaces containing 15,000 and 50,000 synthetic identities per ethnicity with 25 and 45 images per person respectively. These are referred to as z-15k and z-50k for the $\mathcal{Z}$ latent space and w-15k and w-50k for the $\mathcal{W}+$ latent space.

C. Results: Are the Synthetic Identities Biometrically Different?

Since we do not use a facial recognition algorithm in the loop while searching for unique synthetic identities, an important question arises, are the identities biometrically unique? In this section, we experimentally validate this using a SOTA biometric recognition system. The latent space search for synthetic identities is done keeping the perceptual dissimilarity between two consecutive images in mind. We do this by controlling the step size or the range of the uniform random vector. However, we don't provide the search any feedback on the biometric similarity score between two consecutive images that are generated. To validate that these images are in fact biometrically dissimilar, we perform a study using a pre-trained SFace model [31]. The model has been taken from the DeepFace library [47]. We specifically choose SFace as compared to an ArcFace model as it has also been trained on synthetic data and would be better at classifying such data. It achieves similar performance on other metrics compared to the ArcFace model.

We match each person with every other person in the dataset. We show the results in the form of histograms in figure 6 for the *z-50k* and *w-50k* datasets created using the $\mathcal{Z}$ and $\mathcal{W}+$ latent spaces respectively. We use images with the same facial expression, pose, and illumination for every identity to remove any bias from such attributes. In the figure, the red line shows the operating threshold of 0.593 that was set by DeepFace for the SFace model. As visible in the plots, there is an extremely small tail of the histogram that is below the threshold. Implying that the search algorithm can guarantee uniqueness with extremely high accuracy. Additionally, we do not see any low scores, which would have clearly indicated



(a) (b)

Fig. 6. Histogram showing the verification scores on 500,000 generated identities using the (a) $\mathcal{W}+$ and (b) $\mathcal{Z}$ subspace of StyleGAN2. The red line depicts the operating threshold set by DeepFace. As expected there is a larger tail below the threshold for the $\mathcal{Z}$ space-generated identities.

TABLE I
COMPARISON OF TIME IT TAKES (IN MINUTES) TO GENERATE 1000 SAMPLES THROUGH RANDOM REJECTION SAMPLING VERSUS OUR PROPOSED APPROACH FOR DIFFERENT ETHNICITIES. WE SEE OVER 200 TIMES IMPROVEMENT FOR INDIANS AND BLACKS

Demographic Group	Rejection Sampling	Ours
Indian	751.74	31.32
Black	843.84	17.16
White	20.62	13.73
Asian	118.96	17.28
Hispanic Latino	124.31	20.80
Middle Eastern	170.63	22.82

the same person being returned in multiple iterations of the search. This also implies that using this evolutionary search process we are able to generate almost 50 times as many unique biometric identities as there were in the original FFHQ dataset for some ethnicities such as Africans and Indians.

D. Results: Comparison of Computational Time

We empirically compare the computational time required for sampling racially diverse data using random rejection sampling as compared to our proposed approach. In Table I we show results on the time (in minutes) required for both approaches. Random rejection sampling for under-represented groups such as Indians and Blacks is highly inefficient and requires over 12 hours to just generate 1000 samples. Our approach has comparable time for each ethnicity, requiring 32 minutes or less for the same number of samples.

IV. TRAINING THE FACE RECOGNITION MODEL

To show the advantages of the generated racially balanced dataset, we pre-train face recognition models on this dataset. We make use of three facial recognition models, namely ArcFace, ElasticFace, and AdaFace. As a baseline, these models have been trained on a real dataset. We have specifically selected three popularly used datasets in face recognition with varied levels of demographic imbalances - VGGFace2, BUPT-BalancedFace, and BUPT-GlobalFace. We followed the same training and testing protocols as the original authors of the respective recognition models. We provide the training and fine-tuning details including the hyperparameters in Appendix 3.

The VGGFace2 dataset contains 9000+ identities and over 3.3 million images. They do not provide any information on the numbers corresponding to each demographic group.

The BUPT-BalancedFace contains 7000 identities per race but with small variations in the total number of images per race. The subset of the dataset with images of Caucasians contains 326 thousand images and in contrast, the subset for Indians only contains 275 thousand images. Thus even though the dataset is balanced in terms of the number of identities, it has a notable difference in the number of images per identity for the Indian subset.

The BUPT-GlobalFace dataset mimics the demographic distribution that is prevalent in the world. It contains 38 thousand identities and 2 million images in total. 38% percent of the dataset corresponds to white people, 31% to Asians 18% to Indians, and the remaining 13% to Africans.

We report results on the Racial Faces in the Wild (RFW) dataset [48] which contains partitions for 4 racial groups - Caucasians, Blacks, Indians, and Asians. We also report results on the Labeled Faces in the Wild (LFW) [49], Celebrities in Frontal-Profile in the Wild (CFP-FP and CFP-FF) [50], AgeDB [51], Cross-Age LFW (CALFW) [52] and the Cross-Pose LFW (CPLFW) [53] datasets.

The LFW dataset contains 13,233 images of 5,749 people that were extracted using the Viola-Jones face detector algorithm. It is often referred to as the de facto benchmark for unconstrained face recognition. The CFP dataset contains images of celebrities in frontal and profile views. It contains a total of 7,000 pairs of celebrities in both the frontal-frontal (CFP-FF) and frontal-profile (CFP-FP) views. The dataset is primarily used for benchmarking the performance of face recognition across poses. The AgeDB dataset contains 12,240 images of famous personalities, including actors, writers, scientists, and politicians. It contains 440 subjects with varied ages and poses. The dataset is a good benchmark for age-invariant face recognition and age progression. It subject's ages vary from 3 years to 101 years. Similarly, the Cross-Age LFW dataset has been used as a testbed for face recognition across age groups. It has been created from the LFW dataset where 3,000 pairs of images have been selected with age gaps to add aging progression intra-class variance. Similarly negative pairs were also selected with the same gender and race to reduce the influence of other attributes. The CPLFW dataset on the contrary focuses on adding positive subjects with pose variations. They also add 3,000 positive pairs with varied poses and construct the same number of negative pairs keeping the same constraints as the CALFW dataset.

1) Metrics Used for Evaluation: For evaluating the facial recognition models, we report the recognition accuracy on various datasets. Additionally, similar to [54] we utilize the recognition accuracy difference (AD), metric for evaluation of the post-training model biases. Accuracy difference is the maximum difference or disparity between the recognition accuracy of different facets in a set of demographics $\{d_1, d_2, \ldots, d_n\} \in D$ (equation (6)).

$$AD = max_{i,j}|ACC_i - ACC_j|\forall i, j \in D \tag{6}$$

TABLE II
RECOGNITION PERFORMANCE WHEN TRAINING ON THE VGGFACE2 DATASET AND TESTING ON THE SUBSETS OF THE RFW DATASET. THE MODELS TRAINED ON REAL DATA SERVE AS A BASELINE. Z-15K AND Z-50K ARE SYNTHETIC DATASETS CREATED BY SEARCHING ON THE Z-LATENT SPACE WITH 15,000 AND 50,000 IDENTITIES PER RACE RESPECTIVELY. W-15K AND W-50K ARE ONES CREATED USING THE W-LATENT SPACE. W-50K-4 AND Z-50K-4 ARE SUBSETS OF THE SYNTHETIC DATASETS CONTAINING ONLY 4 RACIAL GROUPS SIMILAR TO THE ONES IN THE RFW DATASET - INDIAN, ASIAN, WHITE, AND AFRICAN

Model	Dataset	RFW				AD
		Indian	Asian	White	African	
ArcFace	Real	79.17	74.90	82.48	73.80	8.68
	w-15k	79.43	75.42	82.03	74.58	7.45
	z-15k	77.83	74.77	81.53	74.40	7.13
	w-50k	**80.58**	**77.10**	**83.12**	**76.47**	6.65
	z-50k	80.28	76.92	**83.12**	75.95	7.17
	w-50k-4	78.41	74.80	80.73	74.41	6.32
	z-50k-4	78.15	74.66	79.83	74.25	**5.58**
Adaface	Real	77.97	75.67	82.52	70.18	12.33
	w-15k	77.53	74.50	80.88	70.15	10.73
	z-15k	78.17	75.38	81.30	71.72	**9.58**
	w-50k	77.20	76.07	82.02	68.40	13.62
	z-50k	79.15	77.55	82.73	72.48	10.25
	w-50k-4	77.20	76.35	81.66	70.40	11.26
	z-50k-4	**80.46**	**77.88**	**84.23**	**73.36**	10.87
Elasticface	Real	74.97	71.32	77.78	70.92	**6.87**
	w-15k	79.78	75.87	83.47	75.90	7.57
	z-15k	**81.10**	76.23	**84.62**	**77.08**	7.53
	w-50k	80.92	**76.88**	83.83	76.32	7.52
	z-50k	79.90	75.52	83.72	75.63	8.08
	w-50k-4	79.86	75.63	83.65	74.75	8.90
	z-50k-4	80.23	75.03	84.06	75.36	9.03

TABLE III
RECOGNITION PERFORMANCE WHEN TRAINING ON THE BUPT-BALANCEDFACE DATASET AND TESTING ON THE SUBSETS OF THE RFW DATASET

Model	Dataset	RFW				AD
		Indian	Asian	White	African	
ArcFace	Real	94.23	92.87	95.03	92.92	2.12
	w-15k	94.97	93.70	95.35	93.15	2.20
	z-15k	94.97	93.67	95.25	93.67	1.58
	w-50k	93.73	92.75	94.95	92.38	2.57
	z-50k	94.43	93.00	94.77	92.48	2.28
	w-50k-4	94.18	93.06	95.16	92.43	2.73
	z-50k-4	**95.15**	**94.51**	**95.51**	**93.71**	1.80
Adaface	Real	93.28	92.87	95.02	**90.78**	4.23
	w-15k	**93.43**	92.87	94.97	90.23	4.73
	z-15k	**93.43**	**92.97**	94.62	89.80	4.82
	w-50k	93.33	92.30	94.22	90.18	**4.03**
	z-50k	93.38	92.57	94.37	89.92	4.45
	w-50k-4	92.60	91.98	93.56	89.06	5.50
	z-50k-4	**93.43**	92.66	**95.30**	90.71	4.59
Elasticface	Real	94.23	93.83	95.30	93.03	2.27
	w-15k	94.55	93.98	95.68	93.15	2.53
	z-15k	94.70	93.97	96.02	**93.82**	2.20
	w-50k	94.63	93.50	95.85	93.52	2.33
	z-50k	94.22	93.60	95.77	93.67	**2.10**
	w-50k-4	**95.00**	**94.13**	**96.10**	93.80	2.30
	z-50k-4	94.12	93.58	95.95	93.48	2.47

TABLE IV
RECOGNITION PERFORMANCE WHEN TRAINING ON THE BUPT-GLOBALFACE DATASET AND TESTING ON THE SUBSETS OF THE RFW DATASET

Model	Dataset	RFW				AD
		Indian	Asian	White	African	
ArcFace	Real	94.85	94.28	96.23	93.20	**3.03**
	w-15k	95.10	94.65	**97.27**	92.97	4.30
	z-15k	95.33	95.05	97.00	93.60	3.40
	w-50k	94.98	93.78	96.33	92.60	3.73
	z-50k	95.17	94.13	97.10	92.30	4.80
	w-50k-4	95.07	94.13	96.45	92.75	3.70
	z-50k-4	**95.95**	**95.12**	96.92	**93.87**	3.05
Adaface	Real	94.22	93.88	96.63	91.05	5.58
	w-15k	94.68	**94.22**	96.87	91.52	5.35
	z-15k	**94.80**	94.00	96.57	**91.55**	**5.02**
	w-50k	94.75	93.92	96.72	91.20	5.52
	z-50k	94.72	94.15	**96.97**	91.25	5.72
	w-50k-4	94.67	93.97	96.77	90.93	5.84
	z-50k-4	94.65	93.65	96.85	90.80	6.05
Elasticface	Real	95.32	**94.70**	97.07	93.68	3.38
	w-15k	95.52	94.60	97.28	93.37	3.92
	z-15k	**95.73**	94.38	**97.63**	**93.73**	3.90
	w-50k	94.42	93.28	96.43	91.77	4.67
	z-50k	95.47	94.20	97.03	93.68	**3.35**
	w-50k-4	95.03	94.13	96.83	92.58	4.25
	z-50k-4	94.78	94.25	97.03	92.40	4.63

A. Results on the Use of the Synthetic Dataset for Training Facial Recognition

We hypothesize that using a balanced dataset in terms of ethnic distribution will lead to a more accurate and fair face recognition model. A balanced dataset will help ensure that the model is better able to recognize individuals from underrepresented communities, who may be more likely to be falsely identified by traditional biased models. Most facial recognition models currently are trained on datasets such as MS-1M, and VGGFace2. All of these datasets are unbalanced with respect to ethnic diversity. We compare the advantages of pretraining on the generated synthetic dataset as compared to only training a face recognition model on real data. Given, the distributional shift from high-quality synthetic images to real-world in-the-wild datasets, we finetune all the models trained on balanced synthetic datasets on real-world datasets.

We summarize the results on the RFW dataset in Table II, Table III and Table IV for the VGGFace2, BUPT-BalancedFace and BUPT-GlobalFace datasets respectively. We see a significant improvement in the performance of the models especially in the case of the ElasticFace model when finetuned on the VGGFace2 dataset. The recognition accuracy on the RFW dataset improves from 74.97% to 81.10% for Indians, 71.31% to 76.23% for Asians, 77.78% to 84.62% for Caucasians, and 70.92% to 77.08% for Africans. Even for the AdaFace model, on average, we see an improvement of approximately 2% in the model that was pre-trained with the *z-50k* synthetic dataset.

We also see improvements in the set of experiments involving the BUPT-GlobalFace dataset. However, it is important to note that the RFW dataset has been extracted from the same MS-1M celeb dataset that BUPT-GlobalFace was created from. While the two sets are disjoint, the datasets are similar in terms of the data distribution. Thus, the finetuning on the BUPT-GlobalFace dataset played a larger role in the

TABLE V
RECOGNITION PERFORMANCE WHEN TRAINING ON THE VGGFACE2 DATASET AND TESTING ON THE LFW, CFP-FP, CFP-FF, CALFW, AGEDB, AND THE CPLFW DATASETS

Model	Train Dataset	LFW	CFP-FP	CFP-FF	AGED-DB	CALFW	CPLFW
ArcFace	Real	81.95	57.63	68.51	55.77	90.45	80.60
	w-15k	82.10	56.90	67.83	55.82	90.13	80.70
	z-15k	81.82	58.00	68.90	55.55	90.05	81.33
	w-50k	**82.37**	**58.96**	69.14	55.55	90.82	**81.50**
	z-50k	81.55	57.59	**69.51**	**56.17**	**91.27**	81.00
	w-50k-4	80.96	58.22	69.47	55.81	89.85	79.48
	z-50k-4	81.15	58.00	68.06	55.30	88.66	79.55
Adaface	Real	96.58	85.49	97.80	**84.37**	88.97	79.25
	w-15k	95.57	82.89	97.11	83.33	88.47	77.85
	z-15k	95.98	83.82	97.33	84.30	89.30	78.47
	w-50k	96.75	87.03	97.67	81.63	88.57	79.72
	z-50k	96.68	85.87	98.10	84.22	89.42	79.52
	w-50k-4	95.92	83.57	96.86	83.30	88.73	78.05
	z-50k-4	**97.10**	**87.47**	**98.14**	**85.70**	**89.98**	**80.80**
Elasticface	Real	81.18	58.99	67.49	54.40	87.70	76.90
	w-15k	83.20	61.13	70.36	56.40	91.20	81.52
	z-15k	84.15	60.64	70.96	56.63	91.97	**83.60**
	w-50k	83.05	60.87	71.10	56.88	**92.82**	82.60
	z-50k	83.23	61.16	70.90	56.45	91.02	81.23
	w-50k-4	86.12	62.50	72.29	57.67	91.88	79.67
	z-50k-4	**86.98**	**63.14**	**73.43**	**58.02**	92.57	80.12

TABLE VI
RECOGNITION PERFORMANCE WHEN TRAINING ON THE BUPT-BALANCEDFACE DATASET AND TESTING ON THE LFW, CFP-FP, CFP-FF, CALFW, AGEDB, AND THE CPLFW DATASETS

Model	Dataset	LFW	CFP-FP	CFP-FF	AGE-DB	CALFW	CPLFW
ArcFace	Real	86.48	60.50	**71.26**	**58.27**	94.77	90.62
	w-15k	87.25	59.29	70.50	57.75	95.28	90.43
	z-15k	86.92	**60.87**	70.51	57.34	**95.35**	91.03
	w-50k	86.83	60.04	70.50	57.33	95.18	90.53
	z-50k	86.88	58.66	70.61	57.20	95.03	90.00
	w-50k-4	86.95	60.46	70.51	57.85	94.88	90.40
	z-50k-4	**87.51**	60.84	70.57	57.73	95.25	**91.95**
AdaFace	Real	99.43	88.14	98.70	**91.97**	**94.97**	88.38
	w-15k	99.40	87.07	98.63	91.92	94.83	87.53
	z-15k	99.25	87.69	**98.97**	90.63	94.65	87.72
	w-50k	**99.43**	88.86	98.67	91.62	94.73	88.10
	z-50k	99.25	87.99	98.50	90.68	94.53	87.52
	w-50k-4	99.16	87.24	98.41	90.76	94.38	87.98
	z-50k-4	**99.43**	**89.34**	98.73	91.90	94.85	**88.58**
ElasticFace	Real	87.52	61.30	71.60	58.30	95.13	91.28
	w-15k	88.00	61.41	71.19	58.25	95.13	**91.92**
	z-15k	87.80	**61.77**	**72.41**	58.15	**95.35**	91.63
	w-50k	**88.15**	61.16	71.71	**58.68**	95.08	91.25
	z-50k	87.92	61.73	72.00	57.82	95.15	91.40
	w-50k-4	88.02	60.86	72.17	58.42	95.26	91.63
	z-50k-4	88.05	61.21	71.51	58.23	95.30	91.50

final performance. We believe this is the reason behind the less significant improvement for these sets of experiments. Nonetheless, for the model that was pre-trained on the *z-50k* synthetic dataset, we see an improvement of approximately 0.5% on average across the different subsets of the RFW dataset.

Interestingly, we see improvements even for the BUPT-BalancedFace dataset which is already balanced in terms of ethnic diversity. A recent work [8], showed an improvement in the range of 0.45% to 1% for different ethnicities on the RFW dataset by using synthetic data along with the BUPT-BalancedFace dataset. We show improvements in the range of 0.48% to 1.64% for the ArcFace model pre-trained on the *z-50k-4* dataset. It is also important to note that our approach has other added advantages - we do not collect any real data and make use of an already existing generative model. We can do the synthetic data generation in a zero-shot manner using a simple search-based approach without the requirement for training the StyleGAN from scratch as done by [8].

TABLE VII
RECOGNITION PERFORMANCE WHEN TRAINING ON THE BUPT-GLOBALFACE DATASET AND TESTING ON THE LFW, CFP-FP, CFP-FF, CALFW, AGEDB, AND THE CPLFW DATASETS

Model	Train Dataset	LFW	CFP-FP	CFP-FF	AGED-DB	CALFW	CPLFW
ArcFace	Real	87.72	59.81	**71.36**	**59.25**	95.38	90.43
	w-15k	87.58	58.60	70.66	58.82	95.62	90.50
	z-15k	87.82	**60.79**	70.87	59.08	95.40	**91.28**
	w-50k	87.27	57.51	70.86	58.27	**95.63**	89.52
	z-50k	87.28	58.13	70.87	58.28	95.62	89.65
	w-50k-4	87.53	58.94	70.57	58.82	95.40	90.27
	z-50k-4	**87.88**	58.90	71.29	58.93	95.57	91.08
Adaface	Real	**99.60**	86.23	98.97	93.55	95.40	87.40
	w-15k	99.57	85.50	**99.16**	**93.78**	95.13	87.45
	z-15k	99.38	85.61	99.00	93.62	**95.63**	87.52
	w-50k	99.50	**86.84**	99.07	93.77	95.30	**87.85**
	z-50k	99.47	85.31	99.10	93.27	95.48	87.52
	w-50k-4	99.52	85.57	98.80	93.22	95.48	86.98
	z-50k-4	**99.60**	86.53	99.04	93.17	95.47	87.43
Elasticface	Real	88.80	**61.81**	72.20	**60.10**	95.58	91.73
	w-15k	88.83	61.21	72.34	59.42	95.57	**92.10**
	z-15k	88.97	61.36	**72.71**	59.75	95.42	91.95
	w-50k	87.75	59.61	72.07	58.90	**95.65**	90.55
	z-50k	**89.05**	61.80	71.80	59.57	95.48	91.40
	w-50k-4	88.10	60.44	72.31	58.93	95.58	90.92
	z-50k-4	88.03	60.07	72.23	59.70	95.48	91.23

TABLE VIII
RESULTS FOR ETHNICITY AND GENDER CLASSIFICATION ON THE FAIRFACE AND UTKFACE DATASETS. P(A) IS THE STANDARD DEVIATION OF THE CLASSIFICATION ACCURACY ACROSS THE PROTECTED GROUPS. FAIRGRAPE REFERS TO THE APPROACH PROPOSED BY [55]

Task	Method	All	Male	Female	$\rho(A)$	AD	White	Black	Hisp	E-A	SE-A	Indian	ME	$\rho(A)$	AD
Fairface, Race	Standard	71.92	**71.22**	72.70	**1.04**	**1.48**	**76.92**	84.16	**55.72**	79.06	64.48	75.73	**65.17**	**10.02**	**28.43**
	FairGrape	66.8	65.3	68.6	2.35	3.30	72.2	80.3	47.5	75.8	56.3	70.2	48.6	13.4	31.70
	Ours	**72.08**	70.50	**73.85**	2.36	3.35	76.36	**84.45**	55.11	**79.42**	**65.41**	**77.17**	64.63	10.31	29.34
Fairface, Gender	Standard	94.42	**94.45**	93.86	0.48	0.68	94.16	**90.70**	95.11	95.02	93.83	95.19	95.80	**1.70**	**5.10**
	FairGrape	91.1	91.3	91.0	0.20	0.30	90.4	85.4	92.3	90.1	90.5	91.9	92.8	2.47	7.40
	Ours	**94.53**	94.43	**94.64**	**0.14**	**0.21**	**94.77**	90.48	**95.52**	**94.50**	**94.53**	**96.12**	**95.89**	1.90	5.63
UTKFace, Race	Standard	**92.32**	**91.77**	**92.95**	0.83	1.17	**93.85**	**94.10**	-	**93.67**	-	85.71	-	4.08	8.39
	FairGrape	88.7	88.2	89.3	0.78	1.10	90.6	92.2	-	88.9	-	79.0	-	5.93	13.20
	Ours	92.00	91.61	92.46	**0.60**	**0.85**	93.54	93.23	-	92.81	-	**86.42**	-	**3.39**	**7.11**
UTKFace, Gender	Standard	**94.32**	**93.98**	94.71	0.52	0.73	**95.38**	**95.63**	-	**90.80**	-	93.33	-	**2.21**	**4.82**
	FairGrape	92.2	92.0	92.5	**0.31**	**0.50**	92.7	94.0	-	87.9	-	91.3	-	2.61	6.10
	Ours	93.82	92.79	**95.00**	1.56	2.21	94.46	95.41	-	89.94	-	**93.80**	-	2.40	5.47

We also show that the models pre-trained on the balanced synthetic data help in mitigating the bias in the model. For the Adaface models trained on the VGGFace2 dataset, we see a 2.55% reduction in the maximum disparity between different racial groups. Similarly, there is a 2.02% reduction for the Arcface model. However, there is a slight increase in the accuracy difference for the ElasticFace model. We see similar improvements even in the case of the models trained on BUPT-BalancedFace which is already trained on an unbiased dataset. Thus, the approach boosts the performance of the models while simultaneously reducing the bias.

Finally, we also improvements in the recognition performance for the other datasets - LFW, AgeDB, CFP-FP, CFP-FF, CPLFW, and, CALFW. We reported these results in Table V for the VGGFace dataset, Table VI for the BUPT-BalancedFace dataset and Table VII for the BUPT-GlobalFace dataset. We see the maximum improvements in the case of the AdaFace model when compared to pretraining on *w-50k* dataset. The performance on the CFP-FP dataset improves from 85.49% to 87.03%. We can attribute any improvement in performance improvements in the CFP-FP dataset to the presence of profile view images in the synthetic training set. We made use of the control

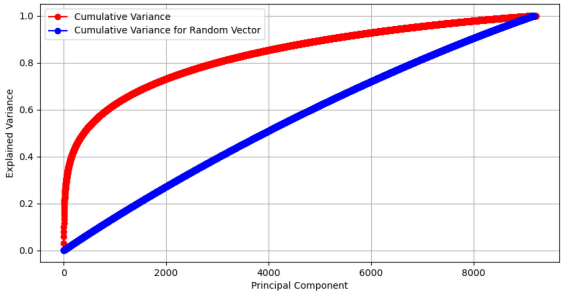


Fig. 7. PCA analysis of 300,000 randomly sampled $\mathcal{W}+$ vectors.

Fig. 8. Examples of 100 different identities corresponding to the "Indian" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression.

over the generative process to include all 180-degree pose variations in the training set.

V. Training the Face Analysis Model

We further show the advantages of synthetic racially balanced data in facial analysis. We specifically focus on two tasks, i.e., ethnicity and gender classification. Similar to FR, the baselines are trained purely on real datasets. We show results on the FairFace [16] and the UTKFace datasets [17].

The FairFace dataset contains 108,501 images that are balanced in terms of racial distribution. The dataset contains 7 racial groups - White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino along with two gender classes Male and Female.

The UTKFace dataset contains over 20,000 images annotated with age, gender, and race. The dataset however is not balanced in terms of racial groups and it contains 4 racial groups - White, Black, Asian, and Indian.

We followed the same training and testing protocols as in [55]. Similar to [55] we reported results on the standard deviation between the performance on the protected groups, referred to as $\rho(A)$. In addition to the accuracy difference metric, this helps us quantitatively access the biases present in the model.

Fig. 9. Examples of 100 different identities corresponding to the "African" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression.

A. Results on Facial Analysis

The results have been summarized in Table VIII. We see that for the FairFace and UTKFace datasets, models pre-trained on our balanced data outperform the approach proposed by [55]. Given the balanced nature of the FairFace and UTKFace datasets, we did not see a significant improvement in the performance over the only training on real data. We expect to see more significant improvements in situations where the real dataset is imbalanced. Nevertheless, for the FairFace dataset we saw that both the accuracy difference (AD) and $\rho(A)$ reduced for gender classification task across the gender groups. However, it was slightly higher for the same task across the racial groups. At the same time, we saw better overall classification accuracy for both race and gender classification.

VI. Discussion

Although GANs offer some control over the data generation process, they also have several limitations. For instance, GANs can only change certain attributes in the variations of each identity, and they cannot replicate real-world data accurately. The generated samples are always consistent in terms of quality and size, which is not the case with real-world data. Consequently, we require additional fine-tuning on a real-world dataset to address this domain gap.

Fig. 10. Examples of 100 different identities corresponding to the "Caucasian" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression.

While we can generate a large number of samples for minority communities we can expect there to be differences in the diversity of samples belonging to the minority communities as there are fewer examples in the original training dataset. Moreover, in the case of generating a combination of protected attributes, this issue is further exacerbated as some combinations may not have been present in the original training dataset. For example, we saw very few examples of 'middle eastern' race and the 'woman' gender group.

In this work, we use a state-of-the-art existing ethnicity, gender, and age classifier [47]. The approach assumes that this classifier is perfect and uses it as supervision for the evolutionary algorithm. We do not consider imperfections in the classifications of the classifier and thus we can expect some noisy predictions or misclassifications. A misclassification can occur in two different situations, the starting latent vector itself has been misclassified and the second case is where a misclassification occurs during a latent space search. In both these cases false positives can introduce some examples of different demographics during the search for a particular demographic group but it would be limited since the classifier would need to constantly misclassify images in a particular latent subspace/ direction to continue the search there. Otherwise, the search would terminate in that latent

Fig. 11. Examples of 100 different identities corresponding to the "Middle Eastern" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression. What is interesting is that most samples are male. There are two possible reasons - biased nature of the ethnicity classifier and the absence of female middle eastern humans in the GAN latent space.

space/ direction after its mutations return negative matches with the target demographic. On the other hand, false negatives can negatively impact the search. These misclassifications become even more pertinent in the case of searching for combinations of demographics where the error multiplies. Here any misclassification in any of the protected attributes leads to termination of the search in that direction/ subspace.

In this study, we have utilized the StyleGAN2 generative model due to its disentangled latent space and its ability to generate high-quality facial images. We believe a similar approach can be applied to any generative model with a disentangled latent space such as Latent Diffusion Models. Diffusion-based models have been shown to generate images with high diversity, however, with higher computational time and cost.

VII. Conclusion

In conclusion, this work presents an approach to generate a balanced number of distinct synthetic identities for different demographic subgroups from a highly biased generative model. We do so in a zero-shot manner without training or finetuning a generative model. We show that this approach works well on the StyleGAN2, and is successful in generating

Fig. 12. Examples of 100 different identities corresponding to the "Asian" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression.

over 50,000 synthetic identities per race. Finally, we show that pretraining a face recognition and analysis models on this dataset boosts the performance of the model. Being a balanced dataset it also assists in mitigating the biases in the model and achieves fairer performance across different demographic groups. This shows that this approach is generalizable and balanced datasets generated using this approach can be used for training any downstream task.

APPENDIX

A. Discussion

We experimented with other approaches similar to past researchers [8], [10], [41] that project real data onto the latent space to get synthetic data. In addition to using privacy-sensitive real data, the projection approach tries to give an exact match between the real identity and projection. While for this task, we are only concerned about an estimated ethnicity match between them. This leads the projection operation to generate unclear or often even demonic faces in an effort to match other unnecessary details such as the background and clothes. Moreover, the projection is more difficult for the underrepresented groups where the variations in the biased generative model are considerably lesser. Additionally, this limits the variations of synthetic data that can be generated to variations or interpolations of the projection of the real data. This would also limit the uniqueness of the identities. Along similar lines as [10], we had also experimented with

Fig. 13. Examples of 100 different identities corresponding to the "Latino Hispanic" racial group generated using the proposed approach. These have been randomly selected from the dataset showing different poses and expression.

randomly generating data and using these as references for these approaches instead of projecting real images on the latent space. However, due to the highly biased nature of StyleGAN2, even after generating over 100,000 samples we had very few samples for the under-represented ethnicities (<2000). This made it computationally expensive in terms of both the time required and storage space. Our proposed approach even without making use of any real or synthetic training data is able to generate a more diverse set of unique identities. This makes it both efficient in terms of time and space as it requires no training data to learn latent directions or interpolations of the data.

B. Disentanglement of the StyleGAN Latent Space

Rahimi et al. [9] suggested limited disentanglement of the StyleGAN3 latent space by visualizing t-SNE plots in two dimensions. We however argue that due to limited correlation between the $\mathcal{W}+$ dimensions, it is inadequate to rely solely on the t-SNE visualization. In Figure 7 we show that even for preserving 80% of the energy you need approximately 4000 dimensions of the data. Thus, there doesn't seem to be strong evidence to suggest that the 9216 dimensional $\mathcal{W}+$ can be accuracy represented on a 2 dimensional plane.

C. Hyper-Parameters for Training and Finetuning Facial Recognition Models

We utilize the following hyperparameters for training the respective face recognition model with a ResNet-50 backbone for all the datasets for consistency. We have used the same parameters for finetuning as well. We had experimented with different learning rates for the synthetic datasets but had found these parameters to be the best performing.

1) AdaFace:

- Batch Size: 512
- Epochs: 26
- Learning rate milestones: 12, 20, 24
- Learning rate: 0.1
- m: 0.4
- h: 0.333
- Low-resolution augmentation probability: 0.2
- Crop augmentation probability: 0.2
- Photometric augmentation probability: 0.2

2) ArcFace:

- Embedding size: 512
- Momentum: 0.9
- Weight Decay: 5e-4
- Batch Size: 128
- Learning rate: 0.02
- Epochs: 20
- Margin list: (1.0, 0.5, 0.0)

3) ElasticFace:

- Epoch: 40
- Batch size: 128
- Learning rate: 0.1
- s: 64.0
- m: 0.5
- std: 0.0175
- Momentum: 0.9
- Warmup: -1
- Weight decay: 5e-4
- Embedding size: 512

REFERENCES

[1] P. Voigt and A. Von dem Bussche, *The EU general data protection regulation (GDPR): A Practical Guide*, vol. 10, 1st ed. Cham, Switzerland: Springer Int. Publ., 2017.

[2] J. McCarthy. "Indian supreme court declares privacy a fundamental right." Accessed: Apr. 11, 2023. [Online]. Available: https://www.npr.org/sections/thetwo-way/2017/08/24/545963181/indian-supreme-court-declares-privacy-a-fundamental-right

[3] L. de la Torre. "A guide to the california consumer privacy act of 2018." SSRN. 2018. [Online]. Available: https://ssrn.com/abstract=3275571

[4] R. Vogt, "Facebook users win class cert. In face scan privacy row." Accessed: Jul. 2, 2022. [Online]. Available: https://www.law360.com/articles/1034143/facebook-users-win-class-cert-in-face-scan-privacy-row

[5] J. Bilyk, "Judge won't short-circuit class action accusing Google photos of breaking IL biometric privacy law." 2017. Accessed: Jul. 19, 2022. [Online]. Available: https://cookcountyrecord.com/stories/511086238-judge-won-t-short-circuit-class-action-accusing-google-photos-of-breaking-il-biometric-privacy-law

[6] V. Monroy, "In the united states district court for the northern district of Illinois eastern division," Shutterfly, Inc., San Jose, CA, USA, document 39, 2017. Accessed: Jul. 10, 2022. [Online]. Available: https://law.justia.com/cases/federal/district-courts/illinois/ilndce/1:2016cv10984/334068/39/

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[8] A. Sevastopolsky, Y. Malkov, N. Durasov, L. Verdoliva, and M. Nießner, "How to boost face recognition with StyleGAN?" 2022, *arXiv:2210.10090*.

[9] P. Rahimi, C. Ecabert, and S. Marcel, "Toward responsible face datasets: Modeling the distribution of a disentangled latent space for sampling face images from demographic groups," 2023, *arXiv:2309.08442*.

[10] P. Melzi et al., "GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations," 2023, *arXiv:2305.19962*.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[12] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18750–18759.

[13] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1578–1587.

[14] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2018, pp. 67–74.

[15] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9322–9331.

[16] K. Karkkainen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1548–1558.

[17] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1–10.

[18] A. Jain, N. Memon, and J. Togelius, "Zero-shot racially balanced dataset generation using an existing biased StyleGAN2," 2023, *arXiv:2305.07710*.

[19] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa, "Anatomizing bias in facial analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 12351–12358.

[20] D. Leslie, "Understanding bias in facial recognition technologies," 2020, *arXiv:2010.07023*.

[21] J. S. Anastasi and M. G. Rhodes, "Evidence for an own-age bias in face recognition," *North Amer. J. Psychol.*, vol. 8, no. 2, pp. 237–252, 2006.

[22] S. Mittal, P. Majumdar, M. Vatsa, and R. Singh, "On bias and fairness in deep learning-based facial analysis," in *Handbook of Statistics*. Amsterdam, The Netherlands: Elsevier, 2023.

[23] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2018, pp. 335–340.

[24] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 18–19.

[25] S. Gong, X. Liu, and A. K. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 3414–3424.

[26] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5310–5319.

[27] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9012–9020.

[28] S. Park, S. Hwang, D. Kim, and H. Byun, "Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2403–2411.

[29] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa, "PASS: Protected attribute suppression system for mitigating bias in face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15087–15096.

[30] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2261–2268.

[31] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "SFace: Privacy-friendly and accurate face recognition using synthetic data," in *Proc. IEEE Int. Joint Conf. Biom. (IJCB)*, 2022, pp. 1–11.
[32] M. Kim, F. Liu, A. Jain, and X. Liu, "DCFace: Synthetic face generation with dual condition diffusion model," 2023, *arXiv:2304.07060*.
[33] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9301–9310.
[34] A. Dabouei, F. Taherkhani, S. Soleymani, J. Dawson, and N. Nasrabadi, "Boosting deep face recognition via disentangling appearance and geometry," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 320–329.
[35] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9243–9252.
[36] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, pp. 5464–5478, 2019.
[37] M. Liu et al., "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3673–3682.
[38] G. Parmar, Y. Li, J. Lu, R. Zhang, J.-Y. Zhu, and K. K. Singh, "Spatially-adaptive multilayer selection for GAN inversion and editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11399–11409.
[39] M.-H. Le and N. Carlsson, "StyleID: Identity disentanglement for anonymizing faces," 2022, *arXiv:2212.13791*.
[40] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face identity disentanglement via latent space mapping," 2020, *arXiv:2005.07728*.
[41] L. Colbois, T. de Freitas Pereira, and S. Marcel, "On the use of automatically generated synthetic image datasets for benchmarking face recognition," in *Proc. IEEE Int. Joint Conf. Biom. (IJCB)*, 2021, pp. 1–8.
[42] Y. Alaluf et al., "Third time's the charm? Image and video editing with StyleGAN3," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 204–220.
[43] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.
[44] V. N. Boddeti, G. Sreekumar, and A. Ross, "On the biometric capacity of generative face models," in *Proc. Int. Joint Conf. Biom. (IJCB)*, 2023, pp. 1–10.
[45] S. I. Serengil and A. Ozpinar, "HyperExtended LightFace: A facial attribute analysis framework," in *Proc. Int. Conf. Eng. Emerg. Technol. (ICEET)*, 2021, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICEET53442.2021.9659697
[46] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.
[47] S. I. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recognition framework," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, 2020, pp. 23–27. [Online]. Available: https://doi.org/10.1109/ASYU50717.2020.9259802
[48] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 692–702.
[49] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life'Images, Detect., Align., Recognit.*, 2008, pp. 1–15.
[50] S. Sengupta, J. Cheng, C. Castillo, V. Patel, R. Chellappa, and D. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Conf. Appl. of Comput. Vis.*, 2016, pp. 1–9.
[51] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, vol. 2, 2017, pp. 1997–2005.
[52] T. Zheng, W. Deng, and J. Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*.
[53] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing Univ. Posts Telecommun., Beijing, China, Rep. 5, Feb. 2018.
[54] A. Jain, N. Memon, and J. Togelius, "A Dataless FaceSwap detection approach using synthetic images," in *Proc. IEEE Int. Joint Conf. Biom. (IJCB)*, 2022, pp. 1–7.
[55] X. Lin, S. Kim, and J. Joo, "FairGRAPE: Fairness-aware gradient pruning method for face attribute classification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 414–432.

Anubhav Jain received the B.Tech. degree (with Hons.) in electronics and communications engineering from the Indraprastha Institute of Information Technology, New Delhi, India. He is currently pursuing the Ph.D. degree with the Tandon School of Engineering, Department of Computer Science and Engineering, New York University. His research interests include computer vision, generative models, and biometrics.

Rishit Dholakia received the bachelor's degree in computer science from the National Institute of Technology Surat, Surat, and the master's degree in computer science from the Courant Institute of Mathematical Sciences, New York University.

Nasir Memon (Fellow, IEEE) received the Bachelor of Engineering degree in chemical engineering and the Master of Science degree in mathematics from the Birla Institute of Technology and Science, Pilani, India, and the Ph.D. degree in computer science from the University of Nebraska. He is a Professor with the Department of Computer Science and Engineering, NYU Tandon School of Engineering, and a Co-Founder of the Center for Cyber-Security, NYU. He has published over 350 articles in journals and conference proceedings and holds a dozen patents in image compression and security. His research interests include digital forensics, biometrics, authentication, security, and human behavior. He has won several awards, including the Jacobs Excellence in Education Award and several best paper awards. He has been on the editorial board of several journals, and was the Editor-in-Chief of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He is a Fellow of IAPR and SPIE.

Julian Togelius received the B.A. degree from Lund University, the M.Sc. degree from the University of Sussex, and the Ph.D. degree from the University of Essex. He has previously worked with IDSIA, Lugano, and with the IT University of Copenhagen. He is a Co-Founder and a Research Director of modl.ai, and an Associate Professor with the Department of Computer Science and Engineering, New York University. He works on artificial intelligence for games and on games for artificial intelligence. His current main research directions involve procedural content generation in games, general video game playing, player modelling, and fair and relevant benchmarking of AI through game-based competitions. Additionally, he works on topics in evolutionary computation, quality-diversity algorithms, and reinforcement learning. From 2018 to 2021, he was the Editor-in-Chief of the IEEE TRANSACTIONS ON GAMES.