Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/177179

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
Abstract—This paper presents a strategy to synthesize face images based on human traits. Specifically, the strategy allows synthesizing face images with similar age, gender, and ethnicity, after discovering groups of people with similar facial features. Our synthesizer is based on unsupervised learning and is capable to generate realistic faces. Our experiments reveal that grouping the training samples according to their similarity can lead to more realistic face images while having semantic control over the synthesis. The proposed strategy achieves competitive performance compared to the state-of-the-art and outperforms the baseline in terms of the Frechet Inception Distance.

Index Terms—Face synthesis, biometrics, unsupervised learning, mixture models

I. INTRODUCTION

Only recently, the synthesis of face images has evolved significantly in terms of quality, albeit with an increase in the complexity of the methods. The state-of-the-art produces high-quality faces with outstanding details [1], [2]. Sythesizing face images can help to train identification models when the training data is scarce or cannot be acquired due to privacy concerns. Moreover, in the context of computer security, this computer vision task is gaining attention to better understand how artificial biometrics are currently generated so they can be timely and accurately identified, for instance, detecting fake social media accounts and preventing identity fraud [3], [4]. Despite the fact that recent methods allow for some basic control over the synthesis process, e.g., generating face images depicting people smiling [5], [6], there are several open aspects to be addressed, for instance, the influence of the training set demographics on the synthesized images. Recent works report an imbalance in the demographic groups depicted in commonly used datasets [7]–[9]. For example, less than a third of the images in the dataset Face Synthetics [9] depict Black, Hispanic, Arab, and Indian individuals, despite the fact that these are the most numerous ethnic groups in the world. This issue requires novel methods for unbiased face synthesis that can preserve the unique facial features of different groups of people while generating face images that result in balanced datasets. In this paper, we introduce a strategy to synthesize face images in an unbiased manner based on different human traits. Our main motivation is to preserve the characteristics of different groups of people during the synthesis process. Our contributions are as follows:

- We show the importance of separately capturing the face features for each group of people for the face synthesis task.
- We achieve state-of-the-art performance in terms of image quality.
- Our strategy can be tailored to any existing face image synthesizer.

The proposed strategy provides the basis for a fairer and unbiased face image synthesis process. Since our strategy preserves the unique facial features of different groups of people, it can help to increase trust in the face image synthesis task. The rest of this paper is organized as follows. Section II summarizes the previous work in face image synthesis. Section III details the proposed strategy. Section IV provides the experiments results and related discussions. Finally, section V concludes this work.

II. PREVIOUS WORK

The methods to synthesize face images can be classified into five groups depending on their methodology. Those based on statistical feature models rely on the data distribution to generate new samples by mapping a noisy training feature space to the synthesized face images. For example, Bordes et al. [10] propose a Markov model to recursively denoise random samples by matching the target distribution from the training dataset. Song et al. [11] propose a Markov model for specific domains to mimic the target data distribution. Ho et al. [12] propose a probabilistic directed graphical model to create a progressive noisy decompressor using variational inference. Prafulla et al. [13] use a CNN filter bank and up-down sampling blocks to generate the images.

The methods based on Generative Adversarial Network (GAN) rely on the discriminator’s capacity to distinguish between the original and the samples synthesized by the generator. For example, Gauthier et al. [5] propose to restrict
the GAN’s generation ability by creating an intermediate space that accepts the noise data along with an embedding to produce an image. Liu et al. [14] propose a two-stream GAN, whose architecture encodes high-level semantics and allows synthesizing pairs of images sharing the same level of abstraction but with a different level of realization. Radford et al. [15] propose a GAN that uses an ad-hoc deep CNN architecture that does not require fully-connected or pooling layers. Yin et al. [16] propose to traverse the latent space using semantic definitions to generate face images with specific characteristics, e.g., smiling or wearing glasses, by using supervised learning. The semantics are also exploited by [17] for face synthesis. Karras et al. propose a coarse-to-fine GAN [8], where the generator and discriminator layers are added as the training progresses. Similarly, Struski et al. [18] constrain the spatial resolution to capture local regions more accurately during the synthesis. Karras et al. [19] further propose to add noise and information from the latent space into the layers’ blocks of the synthesis network in charge of up-sampling the inputs.

Methods based on Variational Auto Encoder (VAE) have recently gained attention because of the high quality of the generated images. Van den Oord et al. [20] propose formulating a VAE in the discrete space via Vector Quantization (VQ), which helps to prevent the posterior collapse issue that occurs when the decoder ignores the latent space. Razavi et al. [21] propose an updated version of the VQ-based VAE, which relies on two deep feed-forward networks. Their method requires two stages, in the first stage a hierarchical VQ-based VAE is trained to encode images into a discrete latent space. In the second stage, a pixel-level CNN is trained to condition the categorical distributions. Rewon et al. [22] show that the VAE requires to be as deep as the data dimension to increase statistical dependence. Although their approach seems very computational expensive, it requires fewer parameters than other VAE-based methods. Vahdat et al. [1] propose a bidirectional encoder comprising residual networks. Similar to the HQ-VAE, their method increases expressiveness in the generated face images by partitioning the latent space.

Methods based on transformers have started to gain importance in computer vision tasks [23], including face image synthesis. Esser et al. [24] 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]. Jiang et al. [25] propose a transformer GAN that is 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.

Finally, methods based on geometric modelling require estimating the face anatomy in order to generate new face images. This process may be manual or synthetic. Cao et al. [26] propose using RGBD cameras to capture an image and subsequently generate a 3D model. Their strategy requires deforming a facial mesh and estimating the captured data with respect to several face landmarks. It then learns the face’s appearance under different deformations. Banerjee et al. [27] propose estimating the face anatomy via Delaunay triangulations, where the face patches are combined and bent to generate new face images. Kim et al. [28] use several facial landmarks to crop face images and train a CNN to estimate the textures. Their model learns pose, shape, expression, skin reflectance, and incident illumination. The authors also propose to learn face features by combining synthetic face images with real ones, a process named breeding. Wood et al. [9] propose a rendering sequence based on hand-crafted facial features to generate face images with outstanding realism and diversity. The synthesis requires polygon masks with several layers of texture.

III. PROPOSED STRATEGY

The proposed strategy is motivated by the importance of preserving and capturing key human traits, i.e., age, gender, and ethnicity, during the face image generation process. Achieving this helps to prevent the generation of imbalance datasets. Since one of our main objectives is to understand
the effect of the training set demographics on the generated images, we select an existing synthesizer based on modesty in terms of computational complexity, i.e., a synthesizer that produces good quality results in an acceptable amount of time. Specifically, we use a GAN-based synthesizer as the backbone of our strategy. Hence, our strategy is not tailored to any particular synthesizer.

Fig. 1 depicts the block diagram of our strategy. It requires splitting the training dataset into several groups and training a synthesizer accordingly to generate the faces for specific groups of people. The strategy requires generating a feature space in an unsupervised fashion. The strategy comprises four stages, which we describe next.

A. Feature Encoding

We use manifold learning to visualize the training dataset and discover groups of people, i.e., clusters, based on their conglomerate distribution. To this end, we first generate the feature space via IR-SE-50 [29], [30], 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. Once the feature space is generated, we learn the manifold using a dimensionality reduction technique.

B. Feature Projection

The feature space is mapped into a low-dimensional space to facilitate discovering the groups. After this mapping, we create a matrix \( X = \{x_1, x_2, \ldots, x_N\} \) containing the low-dimensional samples. The projected data are used to train a mixture model, as detailed next.

C. Group Generation

The matrix \( X = \{x_1, x_2, \ldots, x_N\} \) can be considered as a collection of i.i.d samples from an observable distribution. Let us define a mixture model as follows:

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

where \( \theta = \{\pi, \mu, \Sigma\} \) is the parameter set comprising the model weights, means, and covariances, respectively, for \( K \) components. To estimate \( \theta \), we employ the EM algorithm to estimate the maximum likelihood of the mixture of Gaussians. Because a key objective is to maximize the probability of the observed data, \( X \), we use the component that provides the maximum posterior to create our groups:

\[
k : \arg\max_k p(X, \pi, \mu, \Sigma).
\]

IV. EXPERIMENTS

All experiments use the Flick Faces High Quality (FFHQ) dataset [8]. This dataset comprises 70,000 face images taken from the flickr platform. We use this dataset as it contains samples depicting a diverse range of subjects in terms of age, gender, and ethnicity, with no specific criteria used during data collection. Hence, this dataset is useful to explore the synthesis of different groups of people, unlike other existing datasets, e.g., Celebrities A High Quality (CELEBA-HQ) [7], which comprises face images depicting individuals from a specific

---

1https://github.com/NVlabs/ffhq-dataset
2https://www.flickr.com
3https://github.com/tkarras/progressive_growing_of_gans
The proposed strategy is trained with the full dataset. The results in Table II demonstrate the advantages of generating samples on a per-group basis. Finally, Table III tabulates FID values for two other methods based on a GAN and a VAE. Their reported FID values are for all images without generating them by groups. We can see that the proposed strategy achieves very competitive results in terms of image quality generation.

V. Conclusion

This paper presents a strategy to synthesize face images based on key human traits as discovered by an unsupervised clustering approach. The proposed strategy keeps the features of the face images that are unique to each discovered group and age range (the majority are adults) and posing under highly controlled environments.

We use the model proposed by Karras et al. [19] as the baseline for comparison purposes after tailoring as described in Section III-D. However, the baseline is trained with the full dataset and not on a per-group basis.

A. Embeddings

We first evaluate the IR-SE-50 embeddings produced by using different dimensionality reduction techniques. The purpose of this evaluation is to confirm visually if the feature space forms well-defined clusters, or groups. If this is the case, one can then confirm that it is possible to discover groups of people based on the similarity of their low-dimensional feature vectors. To this end, we test eight different dimensionality reduction methods and randomly visualize 100 samples per method. Specifically, we test Isomaps, Local Linear Embedding (LLE) (three kernels), PCA, hessian eigenmapping, Spectral Embedding (SE), and T-distributed Stochastic Neighbor Embedding (T-SNE) [31].

Fig. 2 shows the resulting embeddings by using different dimensionality reduction methods. We can visually confirm that T-SNE is capable of generating the best-defined groups. We confirm that by using T-SNE, samples associated with the most dissimilar faces, e.g., the elderly and young, tend to be far from those associated with very similar faces. Hence, by using the T-SNE embeddings, it is possible to generate groups that have very similar visual characteristics. Conversely, IsoMaps fail to cluster samples with very similar features, creating effectively sparse groups. PCA effectively generates only one well-defined group comprising samples with very similar facial features. Finally, all of the LLE embeddings produce very sparse groups and thus no visual similarity can be established.

We evaluate the effect of using different embeddings on the quality of the synthesized images, as detailed next.

B. Synthesis

We evaluate the proposed strategy in terms of quality generation and compare its performance against 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 [32].

From Table I, we can see that the T-SNE embeddings indeed produce the best results. T-SNE maps the IR-SE-50 features into a new space where one can see clearly groups sharing similar visual properties. These results are computed as the average over all groups discovered. They mainly reveal the importance of dimensionality reduction in the face image synthesis. Results on a per group basis are tabulated in Table II. From this table, one can see that with the exception of two groups, i.e., Asian Males and Old White Females, the proposed strategy achieves significantly better results than the baseline. More specifically, the proposed strategy archives the lowest FID values for those groups with more training samples. It is worth clarifying that the FID value reported for the baseline is for all images without generating them by groups. The baseline is then trained with the full dataset. The results in Table II demonstrate the advantages of generating samples on a per group basis. Finally, Table III tabulates FID values for two other methods based on a GAN and a VAE. Their reported FID values are for all images without generating them by groups. We can see that the proposed strategy achieves very competitive results in terms of image quality generation.

We also visually inspect the synthesis results to corroborate the profiling generation per group of people. Figure 3 shows samples generated for each of the identified groups. The identified dominant groups in Fig. 3 are Asian males (all ages), Young latin and white females, Old white females, Young white males, Young white and latin children, Young asian females, Adult latin males, and Old white males. One can see that the proposed strategy can accurately generate faces for each group preserving the facial features.
Fig. 3: Sample synthesized image by the proposed strategy using the FFHQ dataset as the training set. Each row displays samples of each one of the dominant groups discovered by the strategy.
of people during training. The results show outstanding performance in terms of image quality, thus corroborating the effectiveness of the proposed strategy. The proposed strategy is a cornerstone in developing fairer face image synthesis methods, as it adequately captures the facial characteristic of different groups of people.

REFERENCES


