MorphGAN: One-Shot Face Synthesis GAN for Detecting Recognition Bias

Nataniel Ruiz[‡1]
nruiz9@bu.edu
Barry-John Theobald[2]
barryjohn_theobald@apple.com
Anurag Ranjan[2]
anuragr@apple.com
Ahmed Hussein Abdelaziz[2]
hussenabdelaziz@apple.com
Nicholas Apostoloff[2]
napostoloff@apple.com

[1] Boston University
Boston
MA, USA

[2] Apple
Cupertino
CA, USA

Abstract

To detect bias in face recognition networks, it can be useful to probe a network under test using samples in which attributes vary in some controlled way. However, capturing a sufficiently large dataset with specific control over the attributes of interest is difficult. In this work, we describe a simulator that applies specific head pose and facial expression adjustments to images of previously unseen people. The simulator first fits a 3D morphable model to a provided image, applies the desired head pose and facial expression controls, then renders the model into an image. Next, a conditional Generative Adversarial Network (GAN) conditioned on the original image and the rendered morphable model is used to produce the image of the original person with the new facial expression and head pose. We call this conditional GAN – MorphGAN.

Images generated using MorphGAN conserve the identity of the person in the original image, and the provided control over head pose and facial expression allows test sets to be created to identify robustness issues of a facial recognition network with respect to pose and expression. Images generated by MorphGAN can also be used to augment training data. We show that augmenting small datasets of faces with new poses and expressions improves the recognition performance by up to 9% depending on the augmentation and data scarcity.

1 Introduction

Training robust face recognition systems [1] requires diverse training data to avoid bias [19]. However, curating large datasets is difficult. A solution is to use realistic images generated by a generative adversarial network (GAN) [12, 28, 34, 47]. However, the ability to explicitly control facial expressions [28, 36, 47] and head-rotations [53] using GANs is limited. 3D morphable models (3DMMs) [17, 40] can be used to render face images. However, these

‡ This work was part of Nataniel Ruiz's internship at Apple.

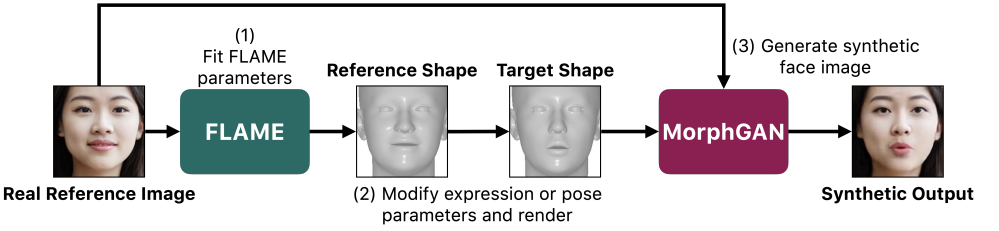

Figure 1: Our face synthesis pipeline: (1) estimate the shape, pose and expression parameters of the FLAME face model, (2) modify the pose and expression parameters, and render the new model instance, and (3) use the rendered model and original image to generate a new image of the same person with the desired pose and facial expression.

images generally lack realism. Ghosh et al. [23] condition a GAN on FLAME [33] to provide explicit controls for head-rotations and facial expressions. Similarly, Tewari et al. [45] condition StyleGAN on the latent representation of a 3DMM. However, in both cases, only faces that arise from the GAN's latent space can be manipulated.

We present an approach that provides controlled manipulation of head pose and facial expression, and preserves identity. Thus, our method can be used to increase the diversity of faces in a dataset to improve the robustness of face recognition models. This is especially important in situations where the number of samples for an individual is limited. We show that the augmentations generated by our model improve face recognition accuracy by up to 9% when limited data are available.

Given a facial image, we first estimate the FLAME parameters to obtain the 3D geometry of the face, and then manipulate the 3D geometry over different head poses and facial expressions. Our conditional GAN takes as input an image and the target 3D geometry, and generates an output image of the person in the target head pose and with the target facial expression. This allows us to perform sensitivity tests of a facial recognition model by generating data samples where only a single attribute is changed [50]. This level of control over the synthesized face allows us to detect and evaluate bias in the facial recognition model.

Contributions. In summary, our main contributions are:

1. *MorphGAN*: a conditional GAN that generates face images with a desired identity, pose and expression by conditioning on a 3D face model and a single reference image of the desired identity.

2. A sensitivity test to detect (pose/expression) bias of a face recognition network using *MorphGAN* generated images of individuals used to train the network.

3. Improved recognition performance using *MorphGAN* images to augment a small training dataset by expanding the variation of pose and expression.

2 Related Work

Early work on face image synthesis used 3DMMs [4], which provide a dense representation of 3D face shape and 2D texture. Furthermore, the model parameters required to represent a specific face can be estimated from an image automatically, and then later used to

manipulate the shape and appearance of the rendered face. Kim et al. [29] used a deep convolutional neural network to refine a rendered 3DMM approximation of a desired image. Geng et al. [22] train networks conditioned on expression coefficients to generate the shape and the texture separately. Ranjan et al. [38] use a convolutional mesh autoencoder to learn a representation of 3D shapes under extreme expressions. More recent works can realistically complete UV maps to render higher quality textured 3DMMs [15]. Some drawbacks of this type of method are that background information is not preserved after re-posing or changing the expression of the face and re-rendering, and that certain face shapes are not accurately captured by 3DMMs, which leads to altered facial features in the final image. Work such as [52] addresses these drawbacks by extracting more detailed face geometry and synthesizing faces on top of existing backgrounds. For this work, a limitation is that continuous fine-grained control of expression is challenging and is left to future work by the authors.

Alternatively, deep generative methods can be applied to the problem of facial image synthesis. For example, X2Face [51] modifies an input face image according to a *driving* source, which could be a different facial image or audio. The approach works well for small changes in head pose or facial expression, but suffers artifacts due to warping if the required transform is large. GANs, and the many variants [24, 26, 27, 48, 49], have received increasing attention because of the high quality images that they can produce. One approach is to condition the network on multiple images of the target person, and input the 2D landmarks of the face in the desired pose and facial expression [53]. However, limitations are a lack of explicit control for pose and expression — the specific landmarks must be provided, and *identity leak* can occur if the landmarks are from a different person. Other work drives changes in the output by imitation of a simulated face [46], although a training set of the identity in question is needed and it is not able to apply these transformations in a one-shot manner. GANimation [37] uses a generator conditioned on action units (AUs) of the Facial Action Coding System (FACS) to generate an attention map to control which areas of the source face need to be modified to transform to a target expression. The advantage of using AUs over landmarks is that AUs provide more intuitive controls than using landmarks. However, a limitation of GANimation is the head pose is fixed. Many recent works in this area achieve high-quality face modifications, although they do so by modifying discrete attributes [16, 42]. In contrast, our work tackles continuous changes in pose and expression. Moreover, recent work that is able to continuously modify pose and expression in faces [16, 42, 45] can do so only on faces that have a corresponding code in the latent space of the face synthesis GAN, whereas our work can modify any face.

The U.S. Department of Commerce released a report showing that contemporary commercial face recognition algorithms exhibit false positive rates that are highest in West and East African and East Asian people, and are lowest in Eastern European individuals [25]. Moreover, gender [6] and other sources of bias [9, 35] have also been detected in face recognition software. One approach to mitigate the effects of bias is to use data augmentation to create samples to re-balance the training data. However, care is required to ensure that mitigating one source of bias does not introduce bias with respect to other attributes [9], including non-obvious sources, e.g. image quality [8]. To identify sources of bias, Kortylewski et al. [31, 32] used a 3DMM to generate data to understand changes in facial recognition rate as a function of a specific facial attribute. However, 3DMMs lack the realism of GAN generated images. An interesting finding in [39] using the *balanced faces in the wild* dataset is that not all data should be considered equal in terms of fairness. For example, the same decision threshold is typically used for all data, which usually hurts certain subgroup(s) even if the overall *best* result is obtained. One particular measure of fairness is parity [8] where

performance is equal across all subgroups. This is a desirable property and can be achieved using a decision threshold that varies by subgroup.

We present a novel GAN-based face image generator with controls that allow precise manipulation of facial attributes. We can use this generator to probe a facial recognition network to understand sources of bias and create samples to re-balance the training data.

3 Methods

Our pipeline for face synthesis, shown in Figure 1, first generates a representation of the target face shape, and then uses this shape representation and a reference image to render a realistic face image. This separation of shape and appearance generation has been demonstrated to be effective [10, 11, 54]. We use FLAME [33] to decompose a face shape into the inherent shape (identity), pose and expression parameters. During training, we use image pairs, a *reference* input and a *target* output. For each image, we extract the 2D facial landmarks and then fit the FLAME model using these landmarks to estimate the FLAME parameters. Our network then learns to generate the *target* image given the *reference* input and *target* FLAME parameters. At test-time, we fit the FLAME parameters to a test image, and then modify the pose and expression parameters, whilst keeping the shape parameters fixed. This updated head model is then used to render the output image of the same person under new head pose and facial expression.

3D Morphable Face Model To control the facial expression and head pose in a photorealistic synthesized image, the generator must be conditioned on both the identity of the person and the desired facial expression and head pose. The representation of expression and pose are important for training the network, and we use a rendered image of a FLAME model [33], as shown in Figure 1. In particular, we generate rendered images, $\boldsymbol{y}$, of the FLAME model, F, that is instantiated using the desired shape, $\boldsymbol{s}$, expression, $\boldsymbol{e}$, and pose, $\boldsymbol{p}$, parameters, and rendered under a certain lighting, ι, using a renderer, R:

$$\boldsymbol{y} = R(F(\boldsymbol{s},\boldsymbol{e},\boldsymbol{p}),\iota). \tag{1}$$

Given an image, we detect the facial landmarks and fit the 3D model to these landmarks to recover $\boldsymbol{s}$, $\boldsymbol{e}$ and $\boldsymbol{p}$. We condition our generator on the original image, and hold ι and $\boldsymbol{s}$ constant, so that new images of the same person can be generated by varying $\boldsymbol{e}$ and $\boldsymbol{p}$.

MorphGAN: Rendering Faces MorphGAN is a conditional image translation GAN, which renders unseen faces in new poses and with new facial expressions. Importantly, MorphGAN requires no fine-tuning or retraining. The generator is conditioned on a reference face image $\boldsymbol{x}_{\text{ref}}$, the rendered target face $\boldsymbol{y}_{\text{tgt}}$ and a style vector ζ_{ref}, as shown in Figure 1.

The MorphGAN generator can be written $G(\boldsymbol{x}_{\text{ref}},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}})$. The style vector ζ is generated by a style encoder network $\zeta = S(\boldsymbol{x})$. We use two discriminator networks, a global discriminator $D_1(\boldsymbol{x},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}})$ and a patch discriminator [26] $D_2(\boldsymbol{x})$, where $\boldsymbol{x}$ is either an image generated by G or a real face image from the training dataset, $\boldsymbol{y}_{\text{tgt}}$ is the rendered face shape, and ζ_{ref} is the style vector extracted from the reference image. We describe the network architectures in the supplementary material.

The networks are trained with losses as described below. We use a supervised perceptual loss $\mathcal{L}_{\text{VGG}}$, using the activations from layers of an ImageNet pre-trained truncated VGG-19

network [43]. We apply L1 loss to the activations of layers $\{4,9\}$ for both the synthesized image and the target image, given by:

$$\mathcal{L}_{\text{VGG}}(\boldsymbol{x}_{\text{ref}},\boldsymbol{x}_{\text{tgt}}) = \sum_{i\in\{4,9\}} M_i||H_1^{(i)}(\boldsymbol{x}_{\text{tgt}}) - H_1^{(i)}(G(\boldsymbol{x}_{\text{ref}},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}}))||_1, \tag{2}$$

where $M_4 = \frac{1}{2}$, $M_9 = 1$, and $H_1^{(i)}$ are the feature outputs for the i-th layer of the VGG-19 network. This perceptual loss helps to capture local detail [20, 21]. To capture higher-level facial features, we use a loss featuring a VGGFace2 pre-trained VGG-13 network [2], where we compute a weighted L1 loss from the activations of convolutional layers $\{10,13\}$ with respective weights for both the synthesized image and the target image. We can write this loss as:

$$\mathcal{L}_{\text{VGGFace}}(\boldsymbol{x}_{\text{ref}},\boldsymbol{x}_{\text{tgt}}) = \sum_{i\in\{10,13\}} W_i||H_2^{(i)}(\boldsymbol{x}_{\text{tgt}}) - H_2^{(i)}(G(\boldsymbol{x}_{\text{ref}},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}}))||_1, \tag{3}$$

where $W_{10} = \frac{1}{2}$ and $W_{13} = 1$, and $H_2^{(i)}$ are the feature outputs for the i-th layer of VGG-13.

We also include a GAN loss $\mathcal{L}_{\text{GAN}}(G,D_1)$ using the global discriminator given by:

$$\mathcal{L}_{\text{GAN}}(G,D_1) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_{\text{tgt}})}[\log D_1(\boldsymbol{x},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}})] + \mathbb{E}[\log(1 - D_1(G(\boldsymbol{x}_{\text{ref}},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}}),\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}})]. \tag{4}$$

Further, the GAN loss $\mathcal{L}_{\text{GAN}}(G,D_2)$, using the patch discriminator is given by

$$\mathcal{L}_{\text{GAN}}(G,D_2) = \mathbb{E}_{\boldsymbol{x}}[\log D_2(\boldsymbol{x})] + \mathbb{E}[\log(1 - D_2(G(\boldsymbol{x}_{\text{ref}},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}}))]. \tag{5}$$

We also include two perceptual cycle consistency losses:

$$\mathcal{L}_{\text{cyc,VGGFace}} = \mathbb{E}_{\boldsymbol{x}}[\mathcal{L}_{\text{VGGFace}}(G(\boldsymbol{x}',\boldsymbol{y}_{\text{ref}},\zeta'),\boldsymbol{x})], \tag{6}$$

$$\mathcal{L}_{\text{cyc,VGG}} = \mathbb{E}_{\boldsymbol{x}}[\mathcal{L}_{\text{VGG}}(G(\boldsymbol{x}',\boldsymbol{y}_{\text{ref}},\zeta'),\boldsymbol{x})], \tag{7}$$

where $\boldsymbol{x}' = G(\boldsymbol{x},\boldsymbol{y}_{\text{tgt}},\zeta_{\text{ref}})$ is the generated sample and $\zeta' = S(\boldsymbol{x}')$ is the style vector predicted from this sample.

Finally, we have two supervised style losses to train the style encoder:

$$\mathcal{L}_{\text{sty,ref}} = \mathbb{E}_{\boldsymbol{x}}[||\zeta' - \zeta_{\text{ref}}||_1], \tag{8}$$

$$\mathcal{L}_{\text{sty,tgt}} = \mathbb{E}_{\boldsymbol{x}}[||\zeta' - \zeta_{\text{tgt}}||_1]. \tag{9}$$

These two losses push the style encoder to generate the same style when the identity of the person is constant.

The final objective function is given by

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{GAN}(G,D_1)} + \mathcal{L}_{\text{GAN}(G,D_2)} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}} + \lambda_{\text{VGGFace}}\mathcal{L}_{\text{VGGFace}} + \\ \lambda_{\text{cyc,VGGFace}}\mathcal{L}_{\text{cyc,VGGFace}} + \lambda_{\text{cyc,VGG}}\mathcal{L}_{\text{cyc,VGG}} + \lambda_{\text{sty,ref}}\mathcal{L}_{\text{sty,ref}} + \lambda_{\text{sty,tgt}}\mathcal{L}_{\text{sty,tgt}}.\end{aligned} \tag{10}$$

Our network architecture and loss design are guided by known best practices to overcome issues observed in generated samples. For the network architecture we use the foundation of [14], with modifications outlined in the Supplementary Material. Perceptual losses are included as they are superior to pixel-level losses for conditional image translation [48, 49, 53]. [14] proposes the supervised style loss, which we use to train the style encoder to preserve identity. [13, 14, 55] use the cycle-consistency loss to transfer between different domains/attributes. Our design choices are informed by this body of work.

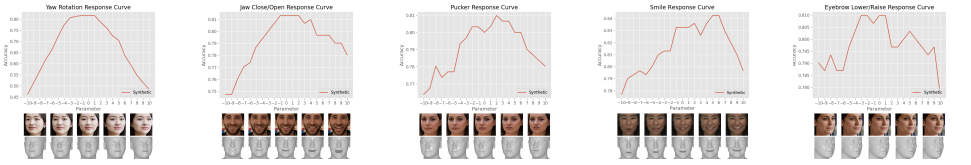

Figure 2: Diagnosis curves for expression and pose changes on a subset of the AVSpeech test data. The examples shown under the curves are generated using MorphGAN and illustrate the specific pose or expression change.

	One Sample		Two Samples		Three Samples	
Parameter	Normal	Augmented	Normal	Augmented	Normal	Augmented
Jaw	71.0%	**77.0%**	86.2%	**89.0%**	90.3%	**90.5%**
Yaw	75.8%	**79.9%**	88.5%	**90.7%**	**91.7%**	91.4%
Smile	69.5%	**76.4%**	86.1%	**88.2%**	89.0%	**90.2%**
Pucker	68.9%	**77.7%**	88.1%	**88.3%**	90.4%	**91.7%**
Eyebrow	73.4%	**73.5%**	86.6%	**86.7%**	89.9%	**91.2%**

Table 1: Face recognition results for a network trained on only real samples (normal) and one trained on a dataset augmented with MorphGAN-generated samples (aug). The top row designates how many real training samples are used to train the network.

4 Experiments

We present three sets of experiments – *diagnosis*, *augmentation* and *face synthesis experiments*. In the diagnosis experiments, we test the sensitivity of a face recognition model by fixing the identity and varying either a pose or expression parameter. In the augmentation experiments, we show improved facial recognition performance after augmenting the training dataset using MorphGAN. Finally, we show qualitative examples of our face synthesis using MorphGAN on unseen samples.

To generate face shape renders, we use the FLAME face model [33]. We train MorphGAN on 13,000 videos from AVSpeech [18]. Our test set consists of a random sample of 150 videos of different people from the AVSpeech test set. We detect the face in each frame and extract the 2D facial landmarks using [6, 30], fit FLAME parameters using the landmarks, and render the resulting shape with a gray texture (see Figure 2 for example renders). In all experiments, our face recognition model is an Inception-ResNet-v1 network [44] pre-trained on the VGGFace2 dataset [7] using the FaceNet triplet loss [41].

4.1 Diagnosis Experiments

We discover robustness issues in a face recognition network by first fine-tuning a pre-trained recognition network using two frames for each of the 150 identities from the AVSpeech test set. Next, we generate new frames using MorphGAN by varying an expression or pose parameter for each of: yaw rotation, jaw opening, smile, lip pucker and eyebrow raise. We generate 21 frames for each parameter, and for each of the images used for fine-tuning.

We discover robustness issues in a face recognition network using images generated by MorphGAN by systematically varying a single expression or pose parameter and measuring the performance of the network as that parameter changes.

We fine-tune the pre-trained recognition network using two frames for each of the 150

Figure 3: Comparison between (1) MorphGAN (ours), (2) unofficial open-source implementation of Zakharov et al. [53] and (3) official implementation of Wiles et al. [51]. The presented expression and pose changes are *(a)* jaw close/open, *(b)* yaw rotation, *(c)* smile and *(d)* pucker. The pink highlighted image is the reference image for each example.

identities from the AVSpeech test set. Next, using MorphGAN we generate new frames by varying the expression or pose for the following parameters: yaw rotation, jaw opening, smile, lip pucker and eyebrow raise. In total, we generate 21 synthetic frames for each of the selected parameters, and for each of the images used for fine-tuning. We choose to explore these expression and pose changes independently to explore specific monocausal sources of bias. In future work, we will look at polycausal interdependent sources of bias, the exploration of which can become arbitrarily complicated as the number and interdependency of the variables increase.

We plot the recognition accuracy for the generated images in Figure 2. Each curve has the ground-truth pose and expression parameter at the center of the x-axis, and varies increasingly from ground-truth value in the negative and positive directions. Also shown are examples of the generated faces and the FLAME model render used to generate each face.

In most cases, the highest accuracy is obtained for the ground-truth pose and expression parameter value. The exceptions are smile and pucker, which peak for a small positive shift in the parameter value. Recognition accuracy drops as the expression and pose shift further from the ground-truth, and yaw rotation impacts accuracy the most (decreases from $> 80\%$ to $\approx 46\%$). Varying the degree of mouth opening, lip pucker and smile also degrades performance (accuracy decreases by up to 7%). In contrast, lowering/raising the eyebrow has only a small impact on accuracy (decrease by $\approx 2\%$) since this is a smaller change in the face than the other parameters. The response curve for some parameters is (approximately) symmetric about the ground-truth, but for others a positive shift is more impactful. Although most curves are strictly increasing in the positive range and strictly decreasing in negative range as expected, there are some exceptions, such as Figure 2.d. We hypothesize that the face recognition network is better at recognizing faces with an expression that is not perfectly neutral. This could be a consequence of training dataset distributions that are not centered around the neutral expression. For example, face recognition datasets, such as CASIA-WebFace, tend to oversample smiling faces since the images are taken from the web. This supports our findings in Figure 2.d. Furthermore, the recognition network is heavily impacted by large changes that cause occlusions, such as yaw rotation, which is similar to findings using real data. We hypothesize that these observations are a result of bias due to the distribution of the samples seen during training.

MorphGAN allows the sensitivity of a network to be tested since parameters can be varied in isolation, and only a single reference image is required. To the best of our knowledge, no other work provides this functionality. Using real data to diagnose bias in this way is impractical since it is impossible to isolate changes caused by only a specific attribute. Fur-

thermore, testing would need to be done using an existing dataset, but all existing datasets have sparse distributions for different expressions and pose parameters — this makes it infeasible to generate continuous curves where a single parameter is varied.

4.2 Augmentation Experiments

It is possible that the generated images impact the face recognition performance. To control for this we generate a *biased* training set and an *unbiased*, uniformly sampled test set for each individual parameter tested in Section 4.1. For the training set, we sample frames that are closest to the mean parameter value of each video, and for the test set we uniformly sample the chosen parameter to create 10 frames. In this way, we build an unbiased uniformly sampled test set and a training set that is biased towards the mean for each parameter. Both of the training and the test sets are composed of only real samples.

In the following experiments, we augment the training data with images generated using MorphGAN and fine-tune the face recognition model on these images. We vary parameters independently to measure their specific effect on recognition accuracy. Contrary to the hypothesis above, when we augment the dataset in this way we show improvement in recognition performance. This suggests that the source of the decrease in performance in the diagnosis experiments (Section 4.1) is the sensitivity of the recognition network for changes in expression and pose, and not due to the images being generated by MorphGAN.

The network is first pre-trained on the large VGGFace2 dataset. We then fine-tune this network under two cases: firstly using only real samples from the AVSpeech dataset, and secondly by augmenting these real samples using MorphGAN to get more variation in head pose or expression. We compare the recognition accuracy in both cases for changes in the same five parameters described in Section 4.1. We also compare the performance when the number of real training samples varies from one to three for each identity.

A *biased* training set, and an *unbiased* uniformly sampled test set are generated for each individual parameter. For the training set, we sample frames that are closest to the mean parameter value of each video, and for the test set we sample 10 frames by uniformly sampling the chosen parameter. In this way, we build an unbiased uniformly sampled test set and a training set that is biased towards the mean for each parameter. Both of these sets are composed of only real samples.

The results for these experiments are shown in Table 1. In the limited-data scenario, where only one training sample per identity is available, we obtain the most gain in performance by augmenting the data set using images from MorphGAN for most of the parameters.

The *increase* in accuracy after augmenting expression and pose is consistent with the *decrease* observed during testing without augmentation. Specifically, the parameters responsible for the largest degradation in accuracy in Section 4.1 (jaw opening and yaw) correspond to the largest increase in accuracy after augmentation. This validates the premise that samples generated by MorphGAN are sufficiently real to be used to probe for sensitivity issues. To reiterate: the model is trained on a dataset of real images augmented with samples generated by MorphGAN, and the model is tested only using real samples.

These results show that MorphGAN can be an effective way of augmenting a dataset for facial recognition, especially in scenarios where there is a limited number of images for each identity, and in which the training data are biased and do not capture the variance with respect to a specific attribute. This experiment also validates the fact that MorphGAN preserves identity, since it improves facial recognition accuracy when testing with only real samples. Finally, we see that we obtain the most significant performance increases (up to

9%) when we augment our training dataset using expression and pose changes that most impacted the facial recognition network in the diagnosis experiment.

4.3 Qualitative Samples

We present qualitative samples of our one-shot face synthesis model to validate identity preservation, expression fidelity and variance, and pose fidelity. In Figure 3, we show comparisons between MorphGAN, an open-source implementation of Zakharov et al. [53] with pre-trained weights, and the official implementation of X2Face [51]. Input images for Zakharov et al. [53] are zoomed out and padded to match their training. In Figure 4, we show examples of expression and pose change on faces generated using the StyleGAN2 architecture trained on the FFHQ dataset. These faces are unseen during MorphGAN training and show the ability of our network to generalize to unseen datasets. We show similarly successful results on other real datasets in the supplementary material.

5 Future Work

In our current implementation, we added terms to the loss function to overcome specific issues observed in generated examples. In future work we will conduct an ablation study to investigate the relative importance of each term. Furthermore, we will extend the idea of investigating bias with respect to expression and pose by applying the same idea to datasets labeled with attributes relating to protected variables, such as ethnicity, gender, and age.

6 Conclusions

In this paper we have presented one-shot generation of images of people with novel facial expressions and head poses using a GAN with interpretable controls. We have used the network to show how bias can be detected in a trained facial recognition network. To synthesize faces we adopt a two step approach: firstly, render the face shape with the desired expression and pose, and secondly render the final image. In particular, we introduce an image generator conditioned on a reference image and the target shape render.

We show that our synthesized face images preserve the identity in the original image, and the synthesized images have high fidelity in expression and pose changes. We have also shown that we can diagnose potential sources of bias with respect to pose and expression. Finally, we have shown that we can improve facial recognition results in a small-data environment by augmenting a source training dataset using additional synthesized face images with new expressions and poses for the corresponding existing identities in the training set.

Acknowledgements

We are grateful to our colleagues Russ Webb and Ashish Shrivastava for their useful feedback.

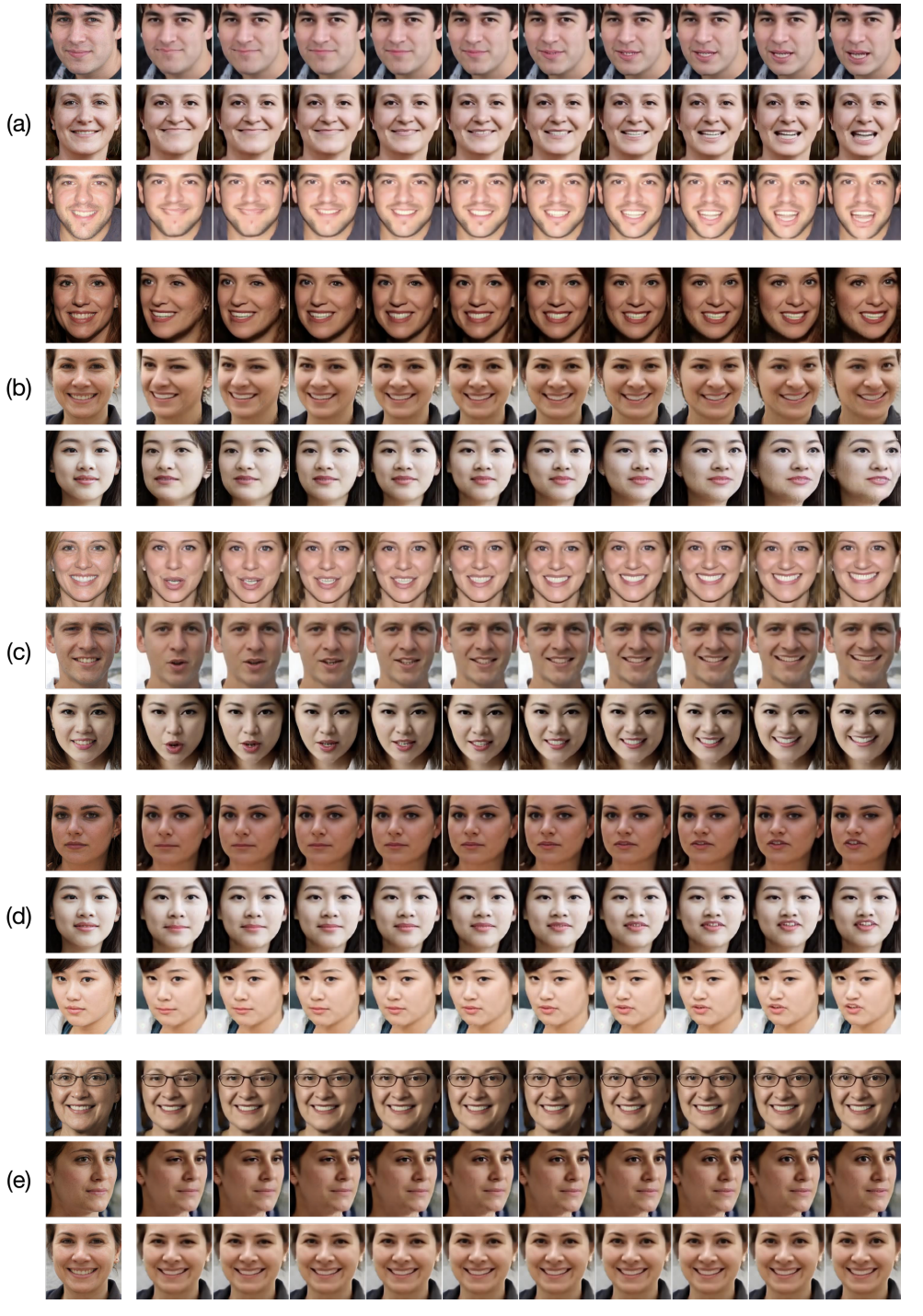

Figure 4: Expression and pose changes using MorphGAN. *(a)* jaw close/open, *(b)* yaw rotation, *(c)* smile, *(d)* pucker and *(e)* eyebrow lower/raise.

References

[1] A. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007.

[2] M. Alvi, A. Zisserman, and C. Nellaaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision Workshops*, September 2018.

[3] R. Bellamy, K. Dey, M. Hind, S. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–15, 2019.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of SIGGRAPH*, pages 187–194, 1999.

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1021–1030, 2017.

[6] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[7] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74. IEEE, 2018.

[8] J. Cavazos, P. Phillips, C. Castillo, and A. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[9] J. Cavazos, P. Phillips, C. Castillo, and A. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[10] L. Chen, R. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[11] Z. Chen, G. Zhang, Z. Zhang, K. Mitchell, and J. Yu. Photo-realistic facial details synthesis from single image. *arXiv preprint arXiv:1903.10873*, 2019.

[12] J. Choe, S. Park, K. Kim, J. Hyun Park, D. Kim, and H. Shim. Face generation for low-shot learning using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1940–1948, 2017.

[13] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[14] Y. Choi, Y. Uh, J. Yoo, and J. Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[15] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018.

[16] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

[17] B. Egger, W. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3D morphable face models — past, present, and future. *ACM Transactions on Graphics*, 39(5):1–38, 2020.

[18] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):112, 2018.

[19] R. Garcia, L. Wandzik, L. Grabner, and J. Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics*, pages 1–6, 2019.

[20] L. Gatys, A. Ecker, and M. Bethge. A neural algorithm of artistic style. In *CoRR*, 2015.

[21] L. Gatys, A. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.

[22] Z. Geng, C. Cao, and S. Tulyakov. 3D guided fine-grained face manipulation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 9821–9830, 2019.

[23] P. Ghosh, P. Gupta, R. Uziel, A. Ranjan, M. Black, and T. Bolkart. GIF: Generative interpretable faces. *arXiv preprint arXiv:2009.00149*, 2020.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[25] P. Grother, Mei L. Ngan, and K. Hanaoka. Face recognition vendor test part 3: Demographic effects, 2019.

[26] P. Isola, J. Zhu, T. Zhou, and A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[27] T. Karras, A. Timo, L. Samuli, and L. Jaakko. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

[28] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[29] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics*, 2018.

[30] D. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[31] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshops*, pages 2093–2102, 2018.

[32] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2261–2268, 2019.

[33] T. Li, T. Bolkart, M. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, November 2017. Two first authors contributed equally.

[34] F. Mokhayeri, K. Kamali, and E. Granger. Cross-domain face synthesis using a controllable GAN. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 252–260, 2020.

[35] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[36] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019.

[37] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision*, pages 818–833, 2018.

[38] A. Ranjan, T. Bolkart, S. Sanyal, and M. Black. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision*, pages 704–720, 2018.

[39] J. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: Too bias, or not too bias? In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshops*, June 2020.

[40] S. Sanyal, T. Bolkart, H. Feng, and M. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, June 2019.

[41] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[42] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[45] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt. Stylerig: Rigging stylegan for 3D control over portrait images. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[46] B. Usman, N. Dufour, K. Saenko, and C. Bregler. Puppetgan: Cross-domain image manipulation by demonstration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9450–9458, 2019.

[47] Y. Viazovetskyi, V. Ivashkin, and E. Kashin. StyleGAN2 distillation for feed-forward image manipulation. *arXiv preprint arXiv:2003.03581*, 2020.

[48] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, pages 1144–1156, 2018.

[49] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[50] B. Webster, S. Kwon, C. Clarizio, S. Anthony, and W. Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *European Conference on Computer Vision*, volume 15, pages 252–270, 2018.

[51] O. Wiles, S. Koepke, and A. Zisserman. X2Face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision*, pages 670–686, 2018.

[52] S. Xu, J. Yang, D. Chen, F. Wen, Y. Deng, Y. Jia, and X. Tong. Deep 3D portrait from a single image. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2020.

[53] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019.

[54] G. Zhenglin, C. Cao, and T. Sergey. 3D guided fine-grained face manipulation. *CoRR*, abs/1902.08900, March 2019.

[55] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017.