Exposing GAN-Generated Profile Photos from Compact Embeddings

Shivansh Mundra[1], Gonzalo J. Aniano Porcile[2], Smit Marvaniya[1], James R. Verbus[2], Hany Farid[2,3]
LinkedIn[1,2] and University of California, Berkeley[3]
Bangalore, India[1], Sunnyvale CA, USA[2], and Berkeley CA, USA[3]
{smundra, ganiano, smarvaniya, jverbus}@linkedin.com and hfarid@berkeley.edu

Abstract

Generative adversarial networks (GANs) have been used to create remarkably realistic images of people. More recently, diffusion-based techniques have taken image synthesis to the next level. From only a text prompt, these techniques can synthesize any image seemingly limited only by our imagination. Along with the many clever and creative use cases, synthetically-generated faces are being used to create more convincing fake social-media profiles. We describe two related techniques that learn low-dimensional (128-D) embeddings of GAN-generated faces. We show that these embeddings capture common facial structures found in these synthetically-generated faces that are uncommon in real profile photos. These low-dimensional models, trained on a relatively small data set, achieve higher classification performance than larger and more complex state-of-the-art classifiers.

1. Introduction

From online dating sites to social media and professional networks, fake profiles, scams, and hoaxes are nothing new. Between January and June of 2019, for example, LinkedIn–at the time, home to more than 645 million members–took action on 21.6 million fake accounts [25]. And, during the first quarter of 2019, Facebook removed 2.2 billion fake profiles[1].

With the rise of GAN-generated synthetic media [14–16] and more recently, text-to-image generated media [1, 23, 24], fake profiles have grown more sophisticated and plentiful. At the same time, the typical user is generally unable to visually distinguish real from synthetically-generated faces [12, 20, 22], and future iterations of synthetic media are likely to contain fewer obvious artifacts.

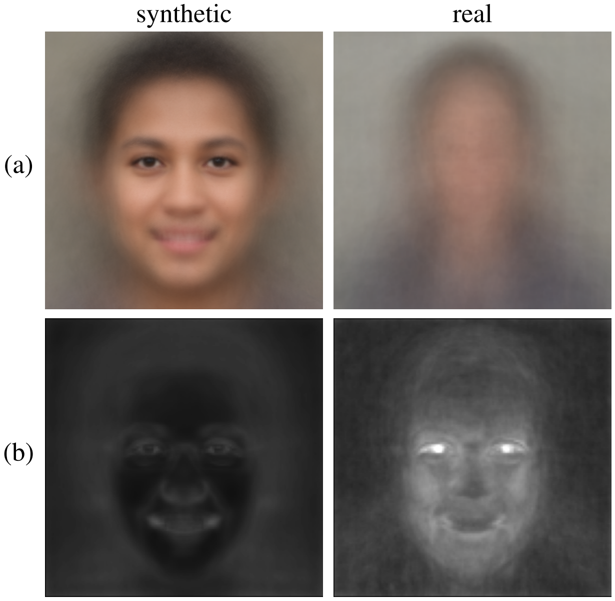

Figure 1. Shown in (a) is the average of 400 StyleGAN2 faces (left) and 400 real profile photos (right), revealing a highly regular synthetic image structure as compared to a highly diverse profile-photo structure. As shown in Figure 2, the StyleGAN2 photos used to create this averaged image were drawn from a diverse demographic pool. Shown in (b) are the average image reconstruction errors (displayed on the same intensity scale) from a learned linear embedding of 10,000 StyleGAN2 faces; this embedding captures the underlying structure of synthetic faces but not profile photos, as seen by the smaller reconstruction error for synthetic faces.

The task of automatically detecting synthetically-generated profile photos is difficult for several reasons: (1) major online platforms are massive: LinkedIn has more than 900 million members[2], with Instagram and Facebook clocking in at 1.2 billion and 3 billion users; (2) there is a significant class imbalance in the prevalence of synthetic and real profile photos, meaning that classifiers with even small error rates of misclassifying real photos as synthetic can be prohibitive when deployed on a major network; (3) the nature of synthetic media is quickly evolving: the past

[1] https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter

[2] https://about.linkedin.com

few years has seen three iterations of increasingly more realistic StyleGAN faces, followed by the most recent text-to-image synthesis; and (4) this is an inherently adversarial system, with the adversary constantly and quickly adapting to new defenses. It is, nevertheless, important that the forensic community develop reliable techniques to distinguish real from synthetic faces that can operate on large networks with hundreds of millions of daily users.

Broadly speaking, there are two categories of forensic approaches to this problem [7]. Hypothesis-driven approaches identify specific artifacts in synthetically-generated faces. For example, inconsistencies in bilateral facial symmetry in the form of corneal reflections and pupil shape have been observed in synthetically-generated faces [10, 11]. Relatedly, head pose and the spatial layout of facial features (eyes, tip of nose, corners of mouth, chin, etc.) in synthetically-generated faces have been observed to be distinct from real faces [30, 31], particularly in earlier incarnations of StyleGAN [15]. The benefit of these approaches is that they learn explicit, semantic-level anomalies. The drawback is that synthesis engines appear to be able to incorporate these features, learning, for example, how to respect bilateral facial symmetry.

Other artifacts include spatial frequency or noise anomalies [3, 9, 21, 33], but these artifacts tend to be vulnerable to simple laundering attacks (e.g., trans-coding and downsampling).

In the second, data-driven category, machine learning is used to learn how to distinguish real from synthetically-generated faces. It has been common for a trained system to accurately classify images from its training set, but then struggle with out-of-domain images. Recently, however, trained systems have shown impressive generalizability across a broad class of synthetic images not included in the training [8, 29]. This generalization appears to be the result of the system learning artifacts introduced as images are resized throughout the image-synthesis pipeline. This strength, however, can also be a weakness because the generic artifacts can also be intentionally or unintentionally removed, rendering these systems useless [5]. Neural-network based approaches are also vulnerable to adversarial attacks where an image is imperceptibly perturbed allowing real images to be easily classified as synthetic and vice versa [2]. The advantage of these techniques, however, is that they can uncover subtle and non-obvious artifacts.

Our approach is a hybrid in which we identify a specific and distinct geometric property in synthetically-generated faces and then use data-driven approaches to quantify and detect these properties. This approach requires training on only a relatively small number of synthesized faces and employs a light-weight classifier that is easy and fast to train.

Shown in the left panel of Figure 1(a) is the average of 400 synthetically-generated (StyleGAN2) faces; shown in the right panel is the average of 400 real (publicly accessible) LinkedIn profile photos. Because the real photos are so varied, the average profile photo is fairly nondescript. In contrast, the average StyleGAN face is highly distinct with almost perfectly focused eyes. This is because StyleGAN faces are aligned in terms of ocular position and interocular distance. In addition to the facial alignment, we also note that StyleGAN faces are primarily synthesized from the neck up, whereas real profile photos tend to show more of the upper body and shoulders. It is this within-class similarity and across-class differences that we seek to exploit.

We describe two related approaches that learn compact embeddings of StyleGAN-generated faces that capture the structural differences illustrated in Figure 1. We then show how these embeddings can be used to distinguish synthetically-generated from real profile photos.

The prior work most related to ours is [17], where the authors use a one-class variational autoencoder (VAE) [18] and a baseline one-class autoencoder [19] to detect deepfake face swaps from the FaceForensics++ dataset [27]. The most significant difference to our work is that we focus on fully synthetic faces (e.g., StyleGAN), while this prior work targets face-swap deepfakes. Second, although we achieve similar overall classification performance, we employ a much simpler and easier to train classifier on a relatively small set of synthetic images.

2. Data Sets

We utilize six data sets consisting of 100,000 real LinkedIn profile photos, and 41,500 synthetically-generated faces spanning five different synthesis engines.

The 100,000 real profile photos were sampled from LinkedIn members with publicly-accessible profile photos uploaded between Jan 1, 2019 and Dec 1, 2022. The accounts used showed activity on the platform on at least 30 days (e.g., signed in, posted, messaged, searched, etc.), without triggering any fake-account detectors.

A total of 10,000 images from each StyleGAN version (1,2,3) [14–16] were downloaded or synthesized. For the first two StyleGAN versions, images were randomly sampled from the larger 100,000 publicly released StyleGAN1[3] and StyleGAN2[4] datasets. For StyleGAN3, we synthesized the images using the released code[5] and pre-trained models[6]. For all three versions, color images were synthesized at a resolution of 1024 × 1024 pixels and with $\psi = 0.5$.[7]

[3] https://github.com/NVlabs/stylegan

[4] https://github.com/NVlabs/stylegan2

[5] https://github.com/NVlabs/stylegan3

[6] We used the stylegan3-r-ffhq-1024x1024.pkl model from the catalog https://catalog.ngc.nvidia.com/orgs/nvidia/teams/research/models/stylegan3.

[7] The StyleGAN parameter ψ (typically in the range $[0, 1]$) controls the truncation of the seed values in the latent space representation used to generate an image. Smaller values of ψ provide better image quality but reduce

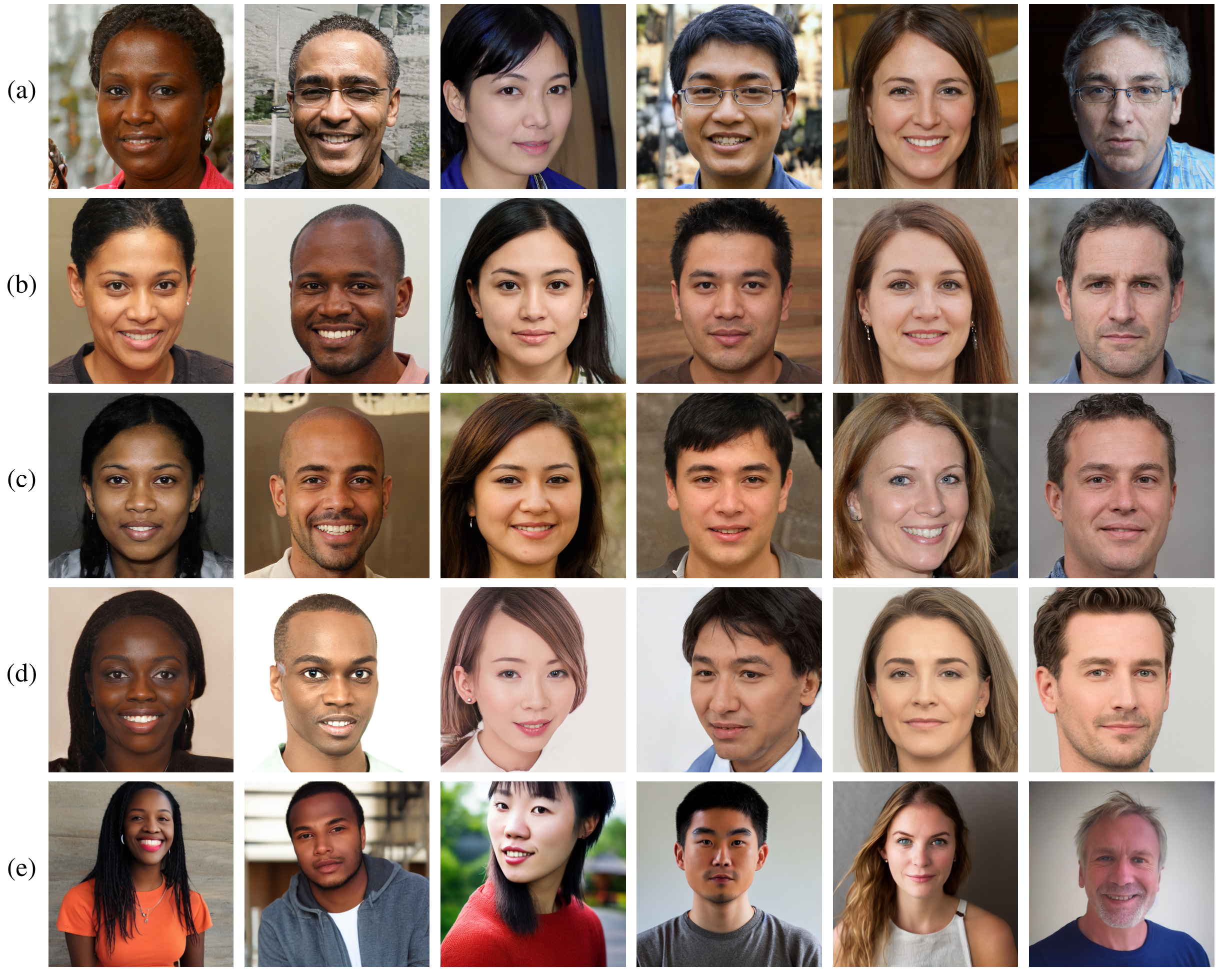

Figure 2. A representative set of synthetic faces from (a) StyleGAN1, (b) StyleGAN2, (c) StyleGAN3, (d) Generated.photos, and (e) Stable Diffusion. In order to respect member privacy, we don't show examples of real profile photos.

A total of 10,000 color images (500 × 500 pixels) were downloaded from Generated.photos[8]. These GAN-synthesized images are generated using a network trained on a proprietary dataset of tens of thousands of high-quality images recorded in a photographic studio.

Lastly, 1500 color images were synthesized using Stable Diffusion [26]. To ensure diversity, fifty faces for each of 30 demographics with the prompts "a profile photo of a {young, middle-aged, older} {black, east-asian, hispanic, south-asian, white} {woman, man}." The color images were synthesized at a resolution of 512 × 512 pixels. This dataset was manually curated to remove obvious synthesis failures in which, for example, the face was not visible.

Shown in Figure 2 are six representative examples from each of these synthetic-generation categories.

All real and synthesized images are subjected to the same pre-processing steps: (1) convert to grayscale; (2) resize to 128×128 pixels in size; and (3) auto-scale into the intensity range $[0, 1]$.

3. Embeddings

As illustrated in Figure 1, we seek a representation (embedding) that captures the spatial alignment common to StyleGAN faces and uncommon to profile photos. This section is partitioned into three parts in which we explore a simple learned linear embedding based on a principal components analysis, a learned embedding based on an autoencoder, and for comparison a fixed linear embedding based on a Fourier analysis. The goal of the latter is to demonstrate that a generic embedding is not sufficient to distinguish synthesized from photographed faces, and that the

facial variety. A mid-range value of $\psi = 0.5$ produces relatively artifact-free faces, while allowing for variation in the gender, age, and ethnicity in the synthesized face.

[8] https://generated.photos/faces

learned embeddings are required to extract sufficiently descriptive representations.

3.1. Learned Linear Embedding

Denote each single-channel, 128 × 128 pixel image as a $128^2 \times 1$ vector $\vec{x}_i$. A principal components analysis (PCA) [6] is used to learn a linear basis from 5000 (one half of the full data set) synthetically-generated faces separately from each of the three StyleGAN synthesis engines (see Section 2). The PCA yields a linear basis where the i^{th} reconstructed image, denoted as $\vec{y}_i$, is:

$$\vec{y}_i = \sum_{j=1}^{m} \alpha_{i,j} \vec{b}_j + \vec{\mu}, \tag{1}$$

where $\vec{b}_j$ are the learned basis vectors, m is the selected basis dimensionality (typically much smaller than 128^2), $\alpha_{i,j} = \vec{x}_i^T \cdot \vec{b}_j$ is the multiplicative contribution of the j^{th} basis, and $\vec{\mu}$ is the mean face across the entire data set subtracted prior to performing the PCA.

With this learned basis, image i is represented as the m-dimensional vector $\vec{\alpha}_i \in \mathcal{R}^m$. We will explore different basis sizes, but initially consider a size of $m = 128$.

We expect that a basis learned from only synthetically-generated faces will accurately capture their structure, while struggling to capture the structure of real profile photos, Figure 1(a). This is quantified using the reconstruction error between an image $\vec{x}_i$ and its low-dimensional reconstruction $\vec{y}_i$, measured as the $\ell 2$-norm: $\|\vec{x}_i - \vec{y}_i\|$.

Shown in Figure 3(a-c) is the normalized distribution of reconstruction errors for 5000 synthetically-generated faces (distinct from those used to construct the linear basis) from each of the three StyleGAN synthesis engines and 100,000 real profile photos.

A specified threshold on this reconstruction error can be used as a simple classifier to distinguish between StyleGAN and profile photos. With this approach, and with a false positive rate[9] (FPR) of 1%, at a threshold of 10.1 across all three classifiers, we observe a true positive rate[10] (TPR) of 71.7% for StyleGAN1; 82.9% for StyleGAN2; and 79.1% for StyleGAN3. Interestingly, we see here that as the StyleGAN faces have improved in photo-realism, they have also become less variable.

Shown in Figure 3(d) is the same distribution in which the PCA was performed on a combination of all three StyleGAN faces. At a threshold of 9.7, we have a TPR of 70.7% for all StyleGAN faces at the same FPR of 1%. Here we see a slight reduction in TPR as compared to individual training.

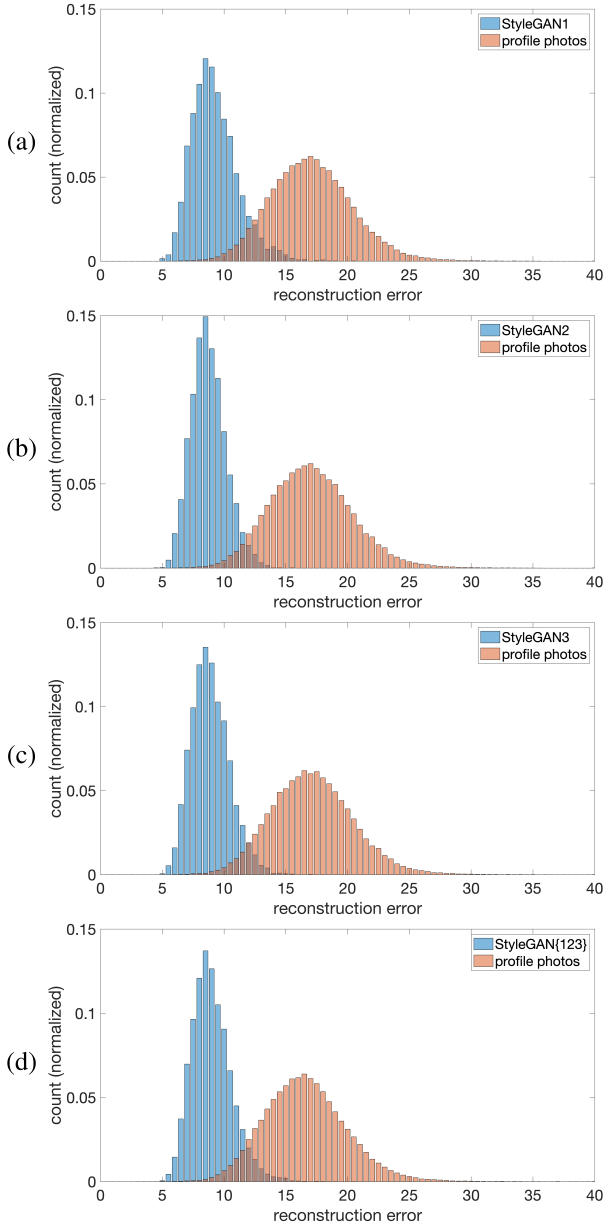

Figure 3. Normalized distributions of image reconstruction error from a learned linear embedding (PCA) (a-c) trained and evaluated separately on three versions of StyleGAN and real profile photos and (d) trained and evaluated on a combination of all three StyleGAN images.

A classifier based on a threshold on reconstruction error is attractive due to its simplicity. It is possible, however, that a classifier based on the underlying low-dimensional embeddings may afford even better discriminatory power.

To this end, we trained a logistic regression on the 128-D training embeddings ($\vec{\alpha}_i$) described above. Trained on the 15,000 combined StyleGAN faces, and with an 80/20 training/testing split, we correctly classify 99.6% of

[9] False positive rate (FPR) is the fraction of real photos that are incorrectly classified as synthetic.

[10] True positive rate (TPR) is the fraction of synthetic photos that are correctly classified as synthetic.

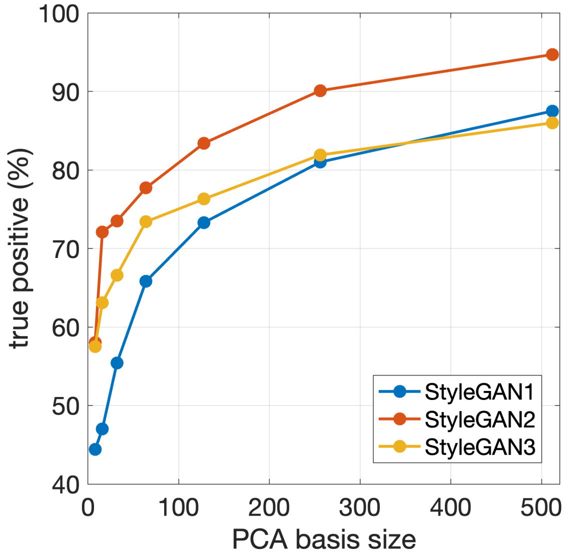

Figure 4. The TPR of correctly classifying a StyleGAN3 synthetic face (with a fixed FPR of 1%) for varying PCA basis size of 8, 16, 32, 64, 128, 256, and 512.

the synthetic faces (TPR) at a 1% FPR. This classifier represents a significant improvement over the above classifier based only on reconstruction error. See Table 1 for a summary of these results.

In the above analysis we, somewhat arbitrarily, fixed the embedding size to $m = 128$. Shown in Figure 4 is the impact of basis size on the TPR with which synthetically-generated faces can be distinguished from real profile photos based only on the reconstruction error. For StyleGAN3 and for a fixed FPR of 1%, the TPR increases steadily from a low of 57.5% for a basis size of $m = 8$ to a high of 86.0% for a basis size of $m = 512$. As compared to a TPR of 79.2% with a basis size of $m = 128$, there is a clear benefit to a larger basis size, but from Figure 4 we see that this benefit begins to plateau after a basis of size 256. This pattern repeats for both StyleGAN1 and StyleGAN2.

3.2. Learned Latent Embedding

With the somewhat surprising efficacy of a simple linear embedding, we next turn our attention to an autoencoder. In particular, we employ a three-layer autoencoder with a first input layer of size 128^2 (the image size), a second hidden layer of size 128 (the same embedding size as that used in the previous section), and a third output layer of size 128^2 (the image size).

The autoencoder is separately trained to reconstruct the 5000 synthetically-generated faces from each of the three StyleGAN synthesis engines (see Section 2). The autoencoder employs a ReLU activation, Adam optimization, an $\ell 2$ regularization term of $\alpha = 0.0001$, and a constant learning rate of 0.001.

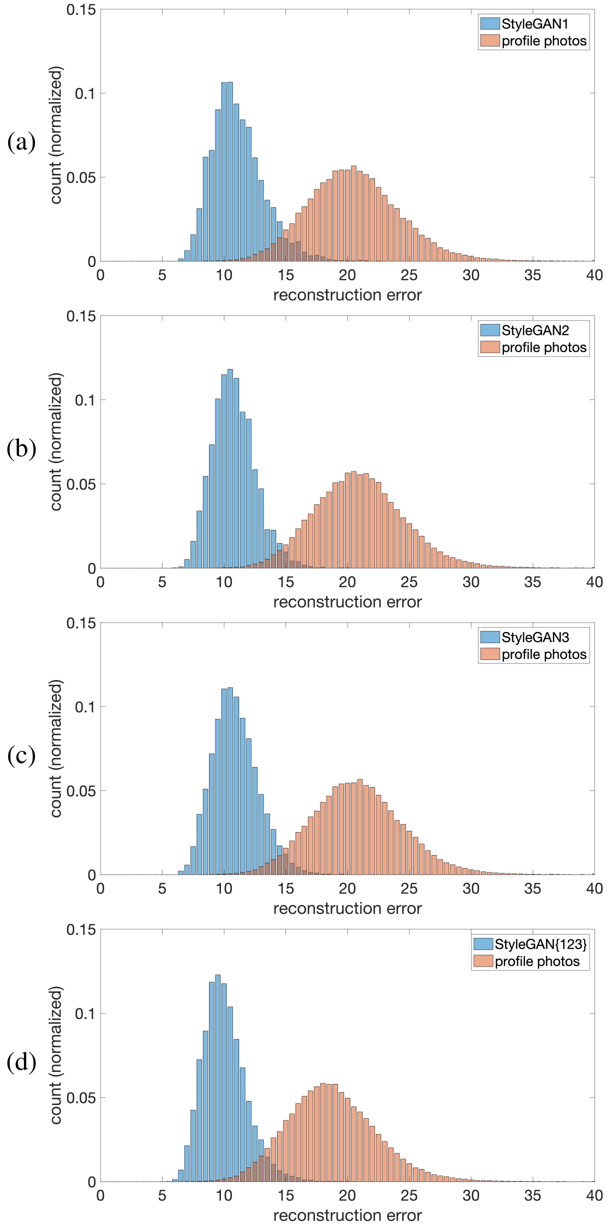

Figure 5. Normalized distributions of image reconstruction error from a learned latent embedding (autoencoder) (a-c) trained and evaluated separately on three versions of StyleGAN and real profile photos and (d) trained and evaluated on a combination of all three StyleGAN images.

As before, we expect that the autoencoder, trained on synthetically-generated faces, will accurately capture their structure, while struggling to capture the structure of real profile photos, Figure 1(a). This difference is again quantified using the reconstruction error between an image $\vec{x}_i$ and its autoencoder reconstruction $\vec{y}_i$, measured as the $\ell 2$-norm: $\|\vec{x}_i - \vec{y}_i\|$.

Shown in Figure 5(a-c) is the normalized distribution of reconstruction errors for 5000 synthetically-generated faces (distinct from those used to construct train the autoencoder) from each of the three StyleGAN synthesis engines and 100,000 real profile photos.

A specified threshold on this reconstruction error can again be used as a simple classifier to distinguish between StyleGAN and profile photos. With this approach, and with an FPR of 1%, at a threshold of 12.7, we observe a 79.7% TPR for StyleGAN1; at a threshold of 13.3, we observe a 92.0% TPR for StyleGAN2; and at a threshold of 12.9, we observe a 86.0% TPR for StyleGAN3.

Shown in Figure 5(d) is the same distribution in which the autoencoder was trained on a combination of all three StyleGAN images. At a threshold of 11.3, we observe a 79.0% TPR for all StyleGAN faces at the same FPR of 1%. Here again we see a slight reduction in TPR as compared to individual training.

The TPR from these learned embedding is, on average, eight percentage points higher than the learned linear embedding (PCA) described in the previous section. See Table 1 for a summary of these results.

As in the previous section, we trained a logistic regression on the 128-D latent representation. Trained on the 15,000 combined StyleGAN faces, and with an 80/20 training/testing split, we correctly classify 99.5% (TPR) of the synthetic faces with a 1% FPR. This classifier represents a significant improvement over the classifier based only on reconstruction error.

3.3. Fixed Linear Embedding

We showed in the previous two sections that a *learned* embedding captures structural similarities of synthetically-generated faces. In this section, we show that a *fixed* representation is unable to capture such differences, emphasizing the importance of the specificity of the learning. Although there are many fixed representations that can be considered, we consider a standard Fourier-based representation.

Denote each 128×128 pixel image as $f_i(x, y)$ and its 2D Fourier transform as $F_i(\omega_x, \omega_y)$. A brick-wall, low-pass filter of size 12×12 is applied to the complex-valued $F_i(\omega_x, \omega_y)$, followed by an inverse Fourier transform to yield a low-dimensional (144-D) representation $g_i(x, y)$. The reconstruction error between the original and reconstruction is measured as the $\ell 2$-norm: $\|f_i(x, y) - g_i(x, y)\|$.

Shown in Figure 6 is the distribution of reconstruction errors for StyleGAN3 and real profile photos. Although the reconstruction errors for the real photos have a longer tail, this fixed embedding does not afford the same discriminability as the learned embedding (Figure 3 and 5): at a threshold of 6.8, only 3.5% of StyleGAN3 faces are correctly classified (TPR) at a 1% FPR. The results are similar for StyleGAN1 and StyleGAN2 are similar.

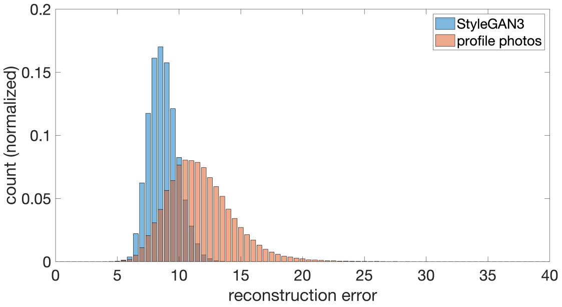

Figure 6. Normalized distributions of image reconstruction error from a fixed linear embedding (Fourier) trained and evaluated on StyleGAN3 and real profile photos. Unlike the learned representations, this fixed representation is unable to distinguish synthetic from real images.

4. Generalization

In the previous sections we showed that low-dimensional embeddings capture structures common to StyleGAN-synthesized faces and uncommon to profile photos. StyleGAN is, of course, only one type of synthetically-generated face. In this section, we analyze the efficacy of our StyleGAN-learned latent embedding (autoencoder) to classify two other categories of synthetically-generated faces (see Section 2.)

Trained on 5000 StyleGAN3 images (see Section 3.2), an autoencoder has a 68.2% TPR for generated.photos images at 1% FPR (real profile photos incorrectly classified as synthetic). By comparison, we observe a 86.0% TPR for StyleGAN3 images. Here we see that the classifier somewhat generalizes, but that the classifier is clearly tuned to specific properties of the StyleGAN training data set.

At the same time, the autoencoder has only a 0.9% TPR for Stable Diffusion images at 1% FPR. This breakdown is not surprising since the diffusion-based process does not rely on the same type of training from aligned faces as the GAN-based process. This breakdown also highlights that our earlier results are, in fact, latching onto a specific property of GAN-generated faces. See Table 1 for a summary of these results.

5. Attacks and Defenses

As described in Section 2, our embeddings are extracted from downsized (to 128×128 pixels) and grayscale converted images. Given this relatively low resolution and the relatively low-dimensional embedding (128-D), it is less likely that our technique will be vulnerable to laundering attacks (resizing, trans-coding, additive noise) or adversarial attacks [2].

Our technique, however, may be vulnerable to simple geometric transformation attacks, which we explore in more

model	classifier	training	testing	TPR
PCA	RE	StyleGAN1	StyleGAN1	71.7%
PCA	RE	StyleGAN2	StyleGAN2	82.9%
PCA	RE	StyleGAN3	StyleGAN3	79.1%
PCA	RE	StyleGAN(123)	StyleGAN(123)	70.7%
PCA	LR	StyleGAN(123)	StyleGAN(123)	99.6%
AE	RE	StyleGAN1	StyleGAN1	79.7%
AE	RE	StyleGAN2	StyleGAN2	92.0%
AE	RE	StyleGAN3	StyleGAN3	86.0%
AE	RE	StyleGAN(123)	StyleGAN(123)	79.0%
AE	LR	StyleGAN(123)	StyleGAN(123)	99.5%
Fourier	RE	StyleGAN3	StyleGAN3	3.5%
AE	RE	StyleGAN3	generated.photos	68.2%
AE	RE	StyleGAN3	Stable Diffusion	0.9%
CNN	CNN	[29]	StyleGAN1	60.1%
CNN	CNN	[29]	StyleGAN2	46.8%
CNN	CNN	[29]	StyleGAN3	9.6%

Table 1. Summary of the rate of correctly identifying synthesized faces (TPR). The model corresponds to principal components analysis (PCA), autoencoder (AE), or Fourier. The classifier corresponds to reconstruction error (RE), logistic regression (LR), or a state-of-the-art CNN [29]. The FPR for the RE and LE is 1%, while the FPR for the CNN is 3.3%.

detail here. To this end, we recreated the PCA basis (Section 3.2) and retrained the autoencoder (Section 3.2) on the StyleGAN3 faces, but this time, randomly cropped and scaled each image. In particular, a central, square bounding box is extracted by randomly stripping between 0 and 12 pixels from each image edge (top, bottom, left, right). The cropped image is then rescaled (using bicubic interpolation) to the original 128 × 128 pixels.

For the recreated PCA basis, a threshold on reconstruction error correctly classifies 22.7% of StyleGAN3 faces (TPR) while incorrectly classifying 1% of real profile photos (FPR). This is a significant reduction in TPR compared to the 79.1% TPR in the absence of cropping and scaling. Similarly, the retrained autoencoder correctly classifies 38.8% of the StyleGAN3 faces (TPR) at the same FPR. This again is a significant reduction in TPR compared to the 86.0% TPR in the absence of cropping and scaling.

As before, a logistic regression trained on the 128-D PCA and autoencoder embeddings yields a significant improvement in classification to a TPR of 77.9% for PCA and 78.8% for autoencoder at the same 1% FPR. While this is lower performance than in the absence of cropping and scaling, our approach is somewhat resilient to this geometric attack.

6. Comparison

The synthetic-image classifier of [29] is a representative example of a state of the art CNN-based image-forensic classifier. This classifier is based on a ResNet-50 architecture pre-trained on ImageNet and then refined to classify an image as photographic or synthesized. The training consists of 720,000 training and 4,000 validation images, half of which are real, the other half of which are ProGAN [13] synthesized images. The training set is augmented with standard image manipulations (e.g., blurring, re-compression). With an average reported precision[11] greater than 90%, the trained classifier can accurately classify ProGAN synthesized images, and, impressively, images from other previously unseen synthesizers.

On our data sets (Section 2), this CNN-based classifier has a TPR of 60.1% for StyleGAN1, 46.8% for StyleGAN2, and only 9.6% for StyleGAN3 faces, with a 3.3% FPR (as compared to our 1% FPR). See Table 1 for a summary of these results.

Despite operating at a 3× higher FPR, the CNN-based classifer TPRs are considerably lower than both our simple reconstruction error and logistic regression based classifiers for all three StyleGAN datsets. At least one reason for this may be that the CNN-based classifier was trained to detect a synthesized image from any category, whereas we focus exclusively on faces. We also see that while the CNN classifier is somewhat able to detect StyleGAN1 and StyleGAN2 images, it struggles significantly on the most recent StyleGAN3 images.

7. Discussion

We have shown that a light-weight, low-dimensional model with relatively minimal training data is highly effective at distinguishing StyleGAN faces from real profile faces. Our approach exploits the fact that all three versions of StyleGAN are trained on cropped and aligned faces, yielding similarly aligned synthetic faces. It remains to be seen, however, if next-generation synthesis engines like 3D-aware GANs [4] or GAN-based text-to-image [28] will exhibit the same facial regularities.

As we were experimenting with the impact of different sized PCA bases on the performance of using only the reconstruction error as a classifier, we observed a monotonic improvement in the TPR at fixed FPR for basis sizes between 8 and 512. Oddly, at a much smaller basis size of 2, 4, and 6, we observed an inversion of this trend with TPRs of 68.8%, 80%, and 70.3%. These TPRs then dipped down to 57.5% for a basis size of 8 before steadily climbing to 86.0% for a basis size of 512, as shown in Figure 4. At the same time, however, a logistic regression trained on these small PCA representations completely fails to accurately distinguish between synthetic and real faces. We posit that with the extra small basis sizes, the PCs latch onto the highly specific eye region (Figure 1). Although these bases yield relatively high reconstruction errors for

[11] Precision is TP / (TP+FP), where true positive (TP) is the number of model predictions where a synthesized image is correctly classified, and false positive (FP) is the number of model predictions where a real image is incorrectly classified as synthesized.

synthetic faces, the reconstruction errors are even higher for real profile photos due to the basis specificity. As the basis size increases to include features other than the eyes, the reconstruction error reduces quickly for both synthetic and real faces. As the basis size increases, the discriminability steadily increases.

The major weakness of our approach is that it is vulnerable to a simple cropping attack. However, because StyleGAN-synthesized images are already fairly tightly cropped around the face, this attack may yield highly atypical profile photos that are visually anomalous. Additional measures may also be applied to counter this attack. In particular, we have employed only the most basic classifiers (principal components analysis, autoencoders, and logistic regression); more sophisticated techniques may be able to learn scale and translation invariant representations.

Whether it was intentional or not, the regularities found in StyleGAN faces are a gift to forensic researchers. Given the potential harms that can come from synthetic media, synthetic-media researchers should consider the addition of distinguishing (but not necessarily perceptible) features as a requirement before broadly deploying their synthesis engines. For example, as described in [32], imperceptible watermarks can be embedded into the training data set after which the synthesis engine learns to generate content with the same watermark. Although not a perfect or full-proof solution, such dataset poisoning would make downstream forensic detection significantly more reliable.

Acknowledgments

This work is the product of a collaboration between Professor Hany Farid and the Trust Data team at LinkedIn[12]. We thank the LinkedIn Scholars[13] program for enabling this collaboration. We also thank Ya Xu, Daniel Olmedilla, Kim Capps-Tanaka, Jenelle Bray, Shaunak Chatterjee, Vidit Jain, Ting Chen, Vipin Gupta, Dinesh Palanivelu, and Milinda Lakkam for their support of this work. We are grateful to David Luebke, Margaret Albrecht, Edwin Nieda, Koki Nagano, George Chellapa, Burak Yoldemir, and Ankit Patel at NVIDIA for facilitating our work by making the StyleGAN generation software, trained models and synthesized images publicly available, and their valuable suggestions.

References

[1] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. arXiv:2103.10951, 2021. 1

[2] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 658–659, 2020. 2, 6

[3] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120, 2020. 2

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *International Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 7

[5] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting GAN-generated fake images from their spectral domain imprints. In *International Conference on Computer Vision and Pattern Recognition*, pages 7865–7874, 2022. 2

[6] George H Dunteman. *Principal Components Analysis*. Sage, 1989. 4

[7] Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022. 2

[8] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. arXiv:2003.08685, 2020. 2

[9] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021. 2

[10] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2904–2908. IEEE, 2022. 2

[11] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2500–2504. IEEE, 2021. 2

[12] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring. Detecting CNN-generated facial images in real-world scenarios. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 642–643, 2020. 1

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196, 2017. 7

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Neural Information Processing Systems*, 2021. 1, 2

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *International Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving

[12] The approach described in this work is not used to take action on any LinkedIn members.

[13] https://careers.linkedin.com/scholars

the image quality of StyleGAN. In *International Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2

[17] Hasam Khalid and Simon S Woo. OC-FakeDect: Classifying deepfakes using one-class variational autoencoder. In *International Conference on Computer Vision and Pattern Recognition*, pages 656–657, 2020. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. arXiv:1312.6114, 2013. 2

[19] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37:233–243, 1991. 2

[20] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces. arXiv: 2106.07226, 2021. 1

[21] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 2

[22] Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[24] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[25] Paul Rockwell. How we're protecting members from fake profiles. https://blog.linkedin.com/2019/august/20/an-update-on-how-were-fighting-fake-accounts, 2019. 1

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *International Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[27] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision*, 2019. 2

[28] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. arXiv:2301.09515, 2023. 7

[29] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *International Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 2, 7

[30] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265. IEEE, 2019. 2

[31] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu. Exposing GAN-synthesized faces using landmark locations. In *ACM Workshop on Information Hiding and Multimedia Security*, pages 113–118, 2019. 2

[32] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *International Conference on Computer Vision and Pattern Recognition*, pages 14448–14457, 2021. 8

[33] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2019. 2