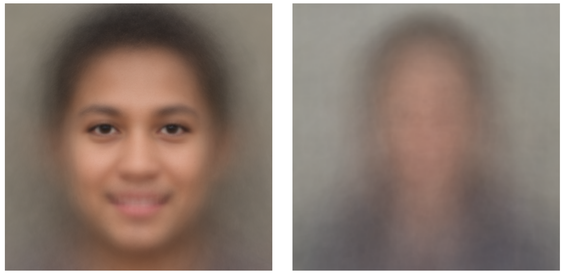
Exposing GAN-Generated Profile Photos from Compact Embeddings

Shivansh Mundra[1], Gonzalo J. Aniano Porcile[1], Smit Marvaniya[1], James R. Verbus[1], Hany Farid[1,2]
LinkedIn[1] and University of California, Berkeley[2]

Overview



Average of 400 GAN (left) and 400 LinkedIn profile images (right) revealing a highly regular GAN facial structure. We learn simple/compact embeddings to capture these GAN regularities.

Dataset

1. StyleGAN1 (10k)
2. StyleGAN2 (10k)
3. StyleGAN3 (10k)
4. generated.photos (10k)
5. stable diffusion (1.5k)
6. *profile photos not shown (100k)*

Methods

Each image is:
1. converted to grayscale
2. resized to 128 × 128 pixels
3. scaled into the intensity range [0,1]

PCA
1. learn a PCA basis
2. project onto 128-D basis
3. compute reconstruction error as $\ell 2$-norm (RE)

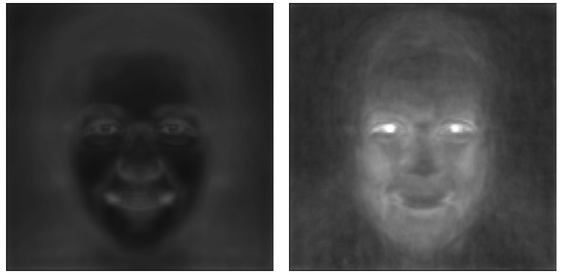
Autoencoder (AE)
1. train an autencoder
2. one hidden layer (n=128)
3. ReLU activation, Adam optimization, $\ell 2$ regularization, constant learning rate
4. compute reconstruction error as $\ell 2$-norm (RE)

Fourier
1. represent using a fixed basis
2. projected onto 12x12 low-pass
3. compute reconstruction error as $\ell 2$-norm (RE)

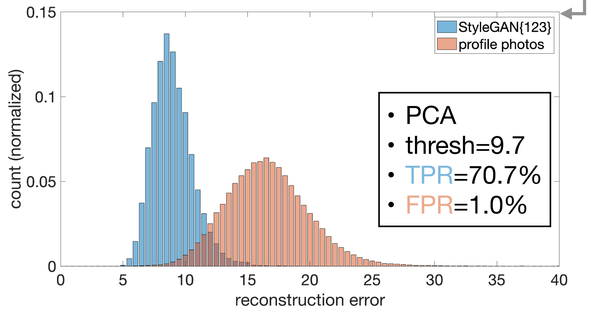
Logistic Regression (LR)
1. train LR classifier on PCA representation
2. train LR classifier on AE representation
3. 80/20 training/testing split (all results on right are for testing)

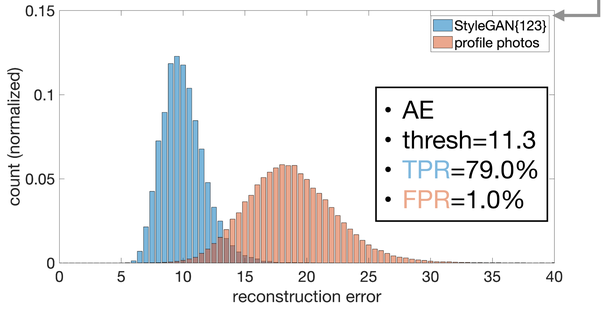
CNN [*]
1. ResNet-50 pre-trained on ImageNet
2. refined to classify an image as real or fake

Results



PCA reconstruction error for GAN (left) and profile (right) [same intensity scale]. See below.





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	[*]	StyleGAN1	60.1%
CNN	CNN	[*]	StyleGAN2	46.8%
CNN	CNN	[*]	StyleGAN3	9.6%

generic Fourier basis fails

generalizes (mostly) to other GAN engines, but doesn't generalize to diffusion synthesis

LR on 128-D outperforms CNN

Robustness to attack
1. retrain PCA/AE on scaled/cropped images
2. TPR for PCA + RE = 22.7%
3. TPR for AE + RE = 38.8%
4. TPR for PCA + LR = 77.9% (down from 99.6%)
5. TPR for AE + LR = 78.8% (down from 99.5%)

[*] Sheng-Yu Wang, et al. CNN-generated images are surprisingly easy to spot... for now. *ICCV*, 2020.