# Finding AI-Generated Faces in the Wild

Gonzalo J. Aniano Porcile[1], Jack Gindi[1], Shivansh Mundra[1], James R. Verbus[1], Hany Farid[1,2]
LinkedIn[1] and University of California, Berkeley[2]
{ganiano,jgindi,smundra,jverbus}@linkedin.com, hfarid@berkeley.edu

## Overview

**We describe a robust and (partially) explainable model for accurate detection of AI-generated faces.**

Operating in the wild and at scale is difficult:

1. major online platforms are massive: LinkedIn has more than 1 billion members;
2. at this scale, even small error rates of misclassifying real photos is prohibitive; and
3. generative-AI is evolving quickly with new techniques often confounding previously trained models.

## Contributions

**Our model has some attractive properties:**

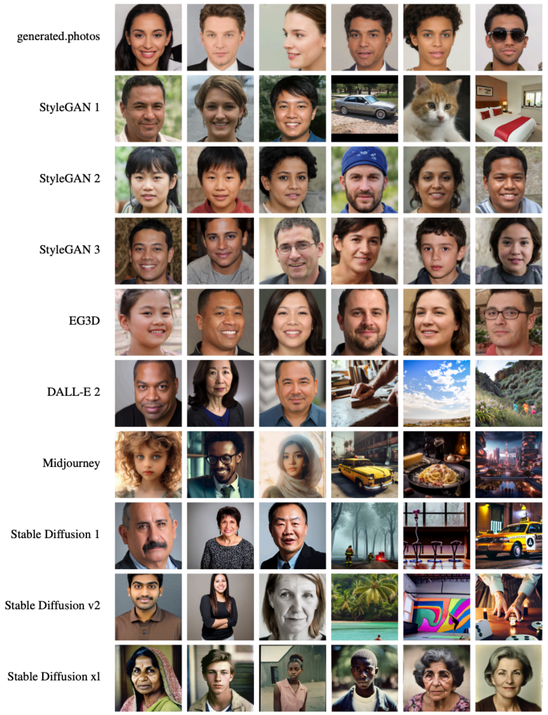
1. it is robust across many generative-AI engines including (in some cases) images from engines not seen during training (see *Dataset*);
2. it operates in the wild on a massive platform; and
3. by focusing on only real/AI faces, we seem to have learned robust, semantic-level features.

*This work describes a model previously operationalized at LinkedIn; this model has since been replaced with a better performing model, allowing us to now talk about this work.*

## Model

**The EfficientNet-B1 base model is trained to distinguish real from AI-generated faces.** This network has 7.8 million internal parameters pre-trained on the ImageNet-1K image dataset.

## Dataset




*Representative examples of >100,000 AI-generated faces (and non-faces). Not shown are examples of 120,000 LinkedIn profile photos.*

## Results

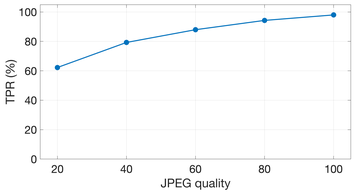
| condition | image | TPR | F1 |
|---|---|---|---|
| training | face | 100.0% | 0.998 |
| evaluation (in-engine) | face | 98.0% | 0.987 |
| evaluation (out-of-engine) | face | 84.5% | 0.914 |
| evaluation (in/out-engine) | non-face | 0.0% | 0.000 |

**evaluation (in-engine)**: With FPR=0.5%, AI faces are correctly classified at 98.0%, varying from 93.3% (Stable Diffusion 1) to 99.5% (StyleGAN 2).

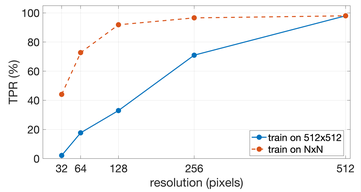
**evaluation (out-of-engine)**: Across synthesis engines not used in training, TPR varied from 19.4% (Midjourney) to 99.5% (EG3D) and 95.4% (generated.photos).

**evaluation (in/out-engine)**: Non-faces (from same synthesis engines used in training) are all classified as real.

**quality**: trained on uncompressed PNG and JPEG images of varying quality, accuracy degrades gently as quality degrades.




**resolution**: trained on 512x512 images, accuracy degrades quickly when images are down- and then up-scaled to 512 (blue). Accuracy improves when trained and evaluated on *NxN* images (red).




## Explainability

10,000 validation images were flipped about horizontal axis and re-classified. With FPR=0.5%, TPR drops from 98.0% to 77.7%.

Flipping about vertical axis has no impact on TPR.

**Combined with robustness to resolution and compression, it appears our model may have latched onto a semantic-level artifact. The analysis below further supports this.**

The unsigned magnitude of the normalized integrated gradients:
(a) StyleGAN 2 (avg. over 100)
(b) DALL-E 2
(c) Midjourney
(d) Stable Diffusion 1
(e) Stable Diffusion 2

The largest gradients are primarily focused on the face and other areas of skin suggesting semantic-level localization.