Synthetic Faces: how perceptually convincing are they?

Sophie J. Nightingale[1,2,*], Shruti Agarwal[2], Erik Härkönen[3], Jaakko Lehtinen[3], and Hany Farid[2]

Lancaster University [1] 2 Berkeley University of California 3 Aalto University

[1,*] s.nightingale1@lancaster.ac.uk

BACKGROUND

Recent advances in machine learning, specifically generative adversarial networks (GANs), have made it possible to synthesize highly photo-realistic faces.

Synthetic faces have been used in the creation of fraudulent social media accounts, including the creation of a fictional candidate for U.S. Congress [1].

In the ongoing fight against misinformation, we examine people's ability to discriminate between synthesized and real faces.

METHODS

Experiment 1. Participants (N=315) were recruited from Mechanical Turk and shown a few examples of synthesized and real faces.

Each participant then saw 128 trials, each consisting of a single face, either synthesized or real, and had unlimited time to classify the face accordingly.

Although unknown to the participant, half of the faces were real and half were synthesized. Across the 128 trials, faces were equally balanced in terms of gender and race.

Experiment 2. A new set of participants (N=170) were recruited from Mechanical Turk. Each participant was given a tutorial consisting of examples of specific artifacts in synthesized faces. Participants were also given feedback after each trial. All other experimental parameters were the same as in Expt. 1.

STIMULI

DATASET

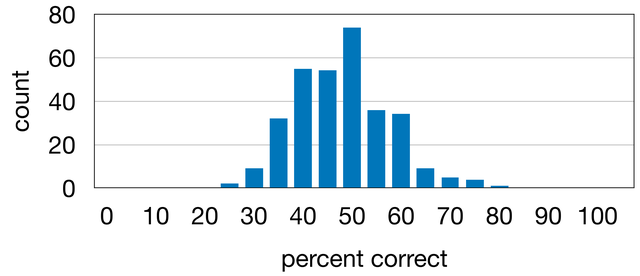
We selected 400 faces synthesized using the state of the art StyleGAN2 [2], ensuring diversity across gender, age, and race.

A convolutional neural network (CNN) descriptor was used to extract a low-dimensional, perceptually meaningful, representation of each face [3].

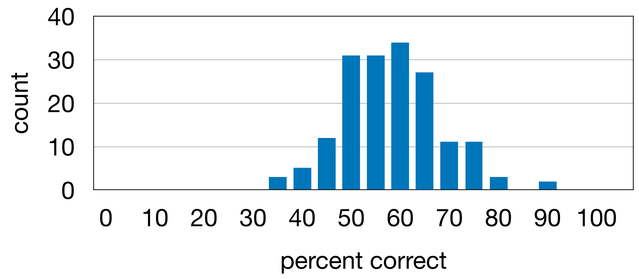
For each of the 400 synthesized faces, this representation was used to find the most similar real faces in the Flickr-Faces-HQ (FFHQ) dataset [4].

RESULTS

Experiment 1. Average performance was close to chance (50%) with no response bias: d' = -0.09; β= 0.99. Below is distribution of individual accuracies.



Experiment 2. Training and feedback improves average performance, but accuracy is still low: d' = 0.41; β= 0.98.



CONCLUSIONS

Images synthesized by StyleGAN2 are realistic enough to fool naive observers.

Even when told about specific synthesis artifacts, observers are unable to reliably discriminate the real from the synthetic.

As synthetic media continues to improve in realism and sophistication, it will become increasingly more difficult to visually discriminate between the real and the fake.

[1] "A high school student created a fake 2020 candidate. Twitter verified it", *CNN*, www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html

[2] T. Karras, et al. "Analyzing and improving the image quality of StyleGAN." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[3] O.M. Parkhi, et al. "Deep face recognition". *Proceedings of the British Machine Vision* Conference, 2015.

[4] https://github.com/NVlabs/ffhq-dataset