Design of Facemasks to Protect from Facial Recognition using Generative Adversarial Networks

Alex Maynard
Department of Aeronautical and Astronautical Engineering
Stanford University
amayn@stanford.edu
https://github.com/alexajm/gan-facemask-design

1 Introduction

Deep learning has empowered substantial improvements in facial recognition technology. While some applications are benign, such as smart phones that unlock when they recognize their user's face, many of the *realized* applications of facial recognition have dangerous implications. For example, while many promote the use of facial id to catch "criminals," in reality it gives the state tremendous power to target minority populations, activists, and other members of our communities. Furthermore, companies such as Clearview AI, who have been criticized for their links to neo-Nazis [1], [2], advertise their facial recognition software as a tool for members of the public to access the private information of anyone they can get a picture of. As such, this project seeks to develop deep learning tools to protect individuals from facial recognition in scenarios when they do not consent to it.

One of the most low-tech but effective tools against facial recognition is a facemask. The advent of the COVID-19 virus and widespread mask use has coincided with some of the most significant political upheaval in the United States in decades. Naturally, this has resulted in research into how to modify facial recognition algorithms to identify people behind their masks. In this project, I propose a network for generating facemask designs that elude such recognition. This is an interesting challenge because it is contrary to much of the existing research on facial recognition, which generally seeks to improve the ability of AI to identify human faces.

2 Relevant Work

Previously, many privacy advocates have used either fashion design or engineering hardware to create surveillance-resilient tools. Examples of fashion-driven solutions include scarves resembling many faces at once [3] or unorthodox makeup [4]. Alternatively, engineering solutions include glasses with infrared LEDs [5] or wearable face projectors that mask the user with a false, projected face [6]. The closest existing work to this project uses convolutional neural networks (CNNs) to generate adversarial "patches" that can be worn around a person's neck to render them invisible to person detection algorithms [7]. This project works at the intersection of fashion design and engineering to develop apparel one could reasonably wear in public, but that still uses adversarial deep learning to protect from facial recognition.

3 Dataset

This project uses images of masked and unmasked individuals from the MaskedFace-Net [8] and Flickr-Faces-HQ (FFHQ) [9] datasets respectively. These datasets provide images of nearly 70,000 individuals to work with, a sample of which are shown in Figure 1.

CS230: Deep Learning, Spring 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Figure 1: Sample of images from the MaskedFace-Net (top) and FFHQ (bottom) datasets.

4 Methodology

I propose a Generative Adversarial Network (GAN) [10] to produce facemask designs that protect the wearer from facial recognition. A vanilla GAN is comprised of two sub-networks: a generator and a discriminator. However, I also employ a third network referred to as the "projector." The generative network samples a latent space and use it to generate a mask design. The projector projects this mask design onto images of masked individuals. These augmented images are then passed to the discriminative network, which returns embeddings that can be used to identify the faces. This is represented visually in Figure 2.

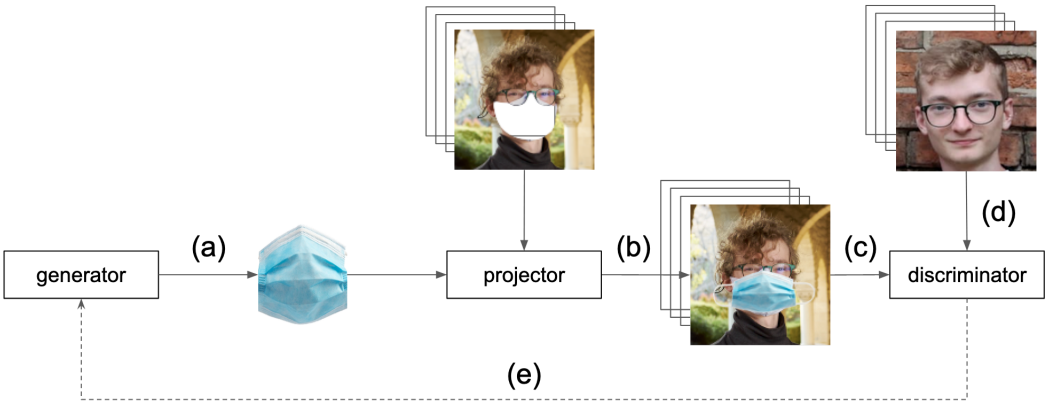

Figure 2: High-level network architecture. (a) generator creates mask design; (b) mask is projected onto training data; (c) masked images are passed to discriminator; (d) discriminator compares masked images against pre-computed embeddings of unmasked images for facial recognition; (e) loss from classification task is used to update generator and discriminator parameters.

4.1 Generator

The generator is modelled as a full-connected 3-layer neural network with 32 units in each layer. It takes in a 100-dimensional vector sampled from $\mathcal{N}(0, 1)$ and outputs a 128x128 RGB image. The generator uses an Adam optimizer, which seeks to *maximize* the cross-entropy loss output by the discriminator. I experimented with other architectures, such as a convolutional neural net (CNN) and a thin, deep, fully-connected neural net, but both were slower and neither showed improved performance over the 3-layer model.

4.2 Projector

The projector has two components: a CNN that removes masks from MaskedFace-Net images and a function that adds a desired mask design in its place. The CNN takes a 128x128 RGB MaskedFace-Net image as input and returns a transparency mask corresponding to the facemask as output. New facemask designs can be projected onto the existing facemask using the output transparency, as shown in Figure 3. The projector is trained prior to the rest of the GAN using a small subset of 100 images from MaskedFace-Net that have had the masks manually replaced with transparent space. Since there is a consistent and predictable pattern in the mask colors and shapes, 100 samples was sufficient to train a projector with low mean squared error loss. After tuning the number of layers and units per layer, I settled on a five-layer network with 6, 12, 18, 24, and 1 channel respectively.

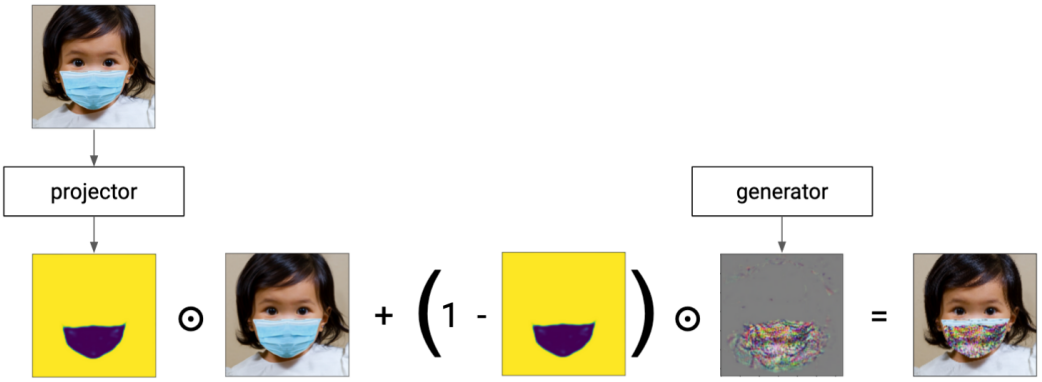

Figure 3: Matrix operations for projecting a generated facemask design onto an existing image using a transparency mask.

4.3 Discriminator

The discriminator is modelled as a standard facial recognition neural network. In particular, I used the `facenet-pytorch` package to load weights for a pre-trained model based on an InceptionV1 network. The network takes in a 128x128 RGB image and outputs a 512-dimensional embedding vector. The embeddings for all unmasked images are pre-computed—to identify a new image, one simply has to find which pre-computed embedding is closest to its own output embedding. All discriminator weights are fixed except for the final 512-unit layer, which is trained alongside the generator. This way, the discriminator learns to recognize faces underneath generated facemask designs, forcing the generator to learn more robust and deceptive designs. This fine-tuning layer is trained to *minimize* a sum of the cross-entropy loss and an L1 regularization term weighted by $\lambda = 0.01$.

5 Experiments

I performed four separate experiments using the GAN described above. First, I fixed all discriminator weights and permitted the GAN to train by itself (experiment 0). I then allowed parameter updates to the final discriminator layer and trained it on

1. generated masks,
2. generated masks and mask-less FFHQ individuals, and
3. generated masks, MaskedFace-Net masks, and mask-less FFHQ individuals.

Each experiment was run for 100 epochs on images of 200 individuals. A collection of designs from all four experiments are shown alongside the original masked images in Figure 4.

The exclusive generator training produces masks that almost appear to mimic human facial features—you can see what appear to be bags under the eyes and horizontal lines that might be

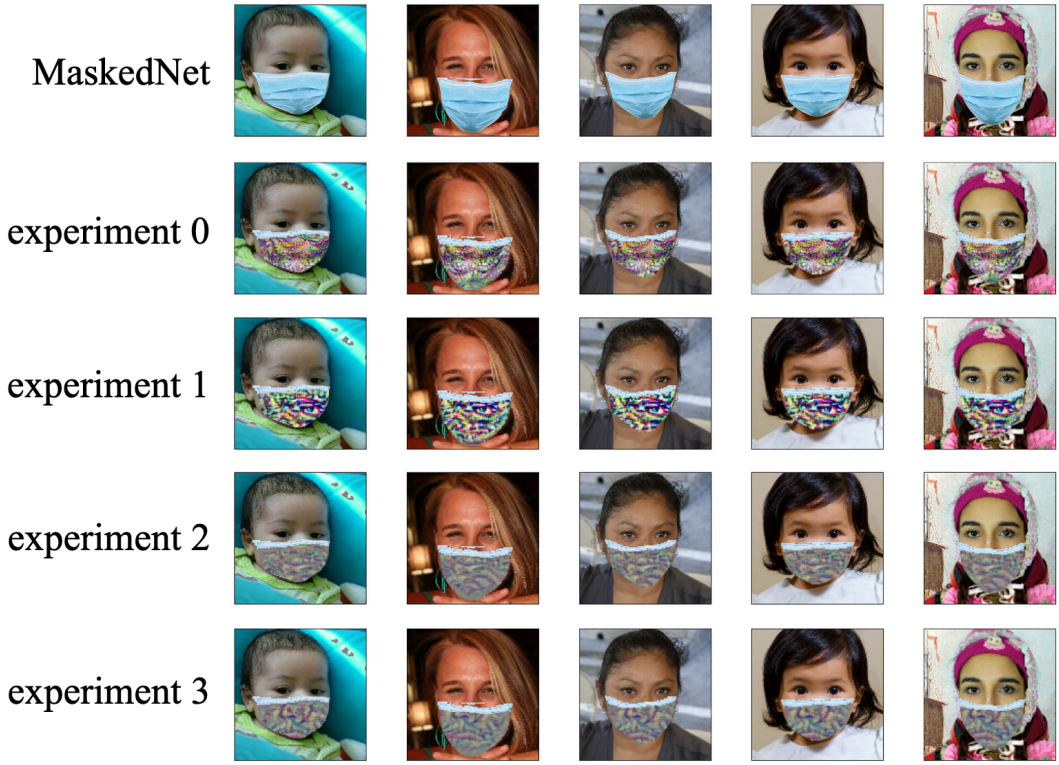

Figure 4: Original MaskedFace-Net images shown alongside designs from (top to bottom) exclusive generator training; discriminator training on generated masks; discriminator training on generated masks and mask-less individuals, and; discriminator training on generated masks, MaskedFace-Net masks, and mask-less individuals.

interpreted as false mouths. When the discriminator started training, the design converged to more of a tie-dye pattern as the generator worked harder to fool the discriminator. Experiment 3 achieved the lowest training accuracy, likely because the discriminator was trying to find a balance between recognizing three different distributions of facial images, preventing it from prioritizing the generator images. Final accuracies for all tests are shown in Table 5, including training accuracy, testing accuracy, and the discriminator's recognition accuracy for the original images from MaskedFace-Net and FFHQ. You may note that the testing accuracy was consistently 0%. This is more likely indicative of a bug than anything else, and upon investigation I noticed that the predicted facial ids were often only a couple indices away from the true facial ids. However, time constraints prevented me from investigating in detail.

experiment #	0	1	2	3
training accuracy	6.2%	15%	23.8%	0.6%
testing accuracy	0%	0%	0%	0%
MaskedFace-Net accuracy	85%	65.1%	73.6%	85.4%
FFHQ accuracy	100%	99%	100%	100%

6 Conclusions and Future Work

Deep learning has empowered large-scale, autonomous facial recognition. While many resulting applications are benign, such tools can be used to infringe on privacies and target marginalized populations. Facemasks are an effective tool against facial recognition, but their widespread use in the past year has incentivized deep facial recognition networks that can identify individuals beneath their masks. This project demonstrated the use of deep neural networks to instead design facemasks that are resilient to facial recognition. In particular, I put a mask-generating network (the generator)

and facial recognition network (the discriminator) in competition in the form of a GAN to discover recognition-resilient facemask designs. The generator was overall very successful, reducing masked recognition from 85% to as low as 0.6% in one case. Furthermore, the fact that the discriminator was simultaneously fine-tuned means these designs could potentially extend to being resilient against more powerful facial recognition algorithms used in the real world. I therefore conclude that deep learning can effectively be used as a tool for protecting individuals from facial recognition.

These results are promising, but there are always improvements to be made. First, the GPU could only work with a couple hundred images at the same time, severely limiting the size of the training dataset. This could feasibly be improved by a memory management system that transfers batches of data between the GPU and CPU as necessary. Second, these masks have been designed digitally given no constraints within RGB-space—it is unclear whether the designs can be practically manufactured, and if so, whether they would function similarly in real life scenarios. An extension of this project might be to print generated masks and test their real-world efficacy. Finally, while the generated masks are sampled from a latent space, they do not vary much within a given model. It would be interesting to experiment with increasing the variance of the generator's mask distribution, or alternatively, conditioning the generator designs on the faces of the people they are intended for.

References

[1] K. Hill, "The Secretive Company That Might End Privacy as We Know It," en-US, *The New York Times*, Jan. 2020, ISSN: 0362-4331. [Online]. Available: https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (visited on 04/21/2021).

[2] L. O'Brien, *Far-Right Extremists Helped Create The World's Most Powerful Facial Recognition Technology*, en, Section: Politics, 400. [Online]. Available: https://www.huffpost.com/entry/clearview-ai-facial-recognition-alt-right_n_5e7d028bc5b6cb08a92a5c48 (visited on 04/21/2021).

[3] S. Weekers, *Anonymous | Sanne Weekers*, en-GB. [Online]. Available: http://sanneweekers.nl/big-brother-is-watching-you/ (visited on 05/17/2021).

[4] A. Harvey, *CV Dazzle: Computer Vision Dazzle Camouflage*. [Online]. Available: https://cvdazzle.com/ (visited on 05/17/2021).

[5] I. Echizen, *Privacy Protection Techniques Using Differences in Human and Device Sensitivity*.

[6] J.-C. Liu, *WEARABLE FACE PROJECTOR*, nl. [Online]. Available: http://jingcailiu.com/wearable-face-projector/ (visited on 05/17/2021).

[7] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," en, *arXiv:1904.08653 [cs]*, Apr. 2019, arXiv: 1904.08653. [Online]. Available: http://arxiv.org/abs/1904.08653 (visited on 05/17/2021).

[8] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "Maskedface-net – a dataset of correctly/incorrectly masked face images in the context of covid-19," *Smart Health*, 2020, ISSN: 2352-6483. DOI: https://doi.org/10.1016/j.smhl.2020.100144. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352648320300362.

[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *CoRR*, vol. abs/1812.04948, 2018. arXiv: 1812.04948. [Online]. Available: http://arxiv.org/abs/1812.04948.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, Jun. 2014, arXiv: 1406.2661. [Online]. Available: http://arxiv.org/abs/1406.2661 (visited on 04/21/2021).