James Kunstle

Privacy-Via-GAN Research Project Proposal

**Statement of Inquiry:**

Can modern GANs, specifically CycleGANs (CGANS), be used to disturb the latent structure of images posted to social media in such a way that they are adversarial to individual recognition? Can this be done in such a way that the image is maximally disturbed from the perspective of recognition networks while remaining minimally changed from the perspective of users?

**Context:**

In the popular VOX article “**Some AI just shouldn’t exist**,” author Sigal Samuel portrays the current state of facial recognition and individual classification DL models as biased and harmful to racial and sexual equality. Indeed, Zoé Samudzi of The Daily Beast argues poignantly:

“*In a country where crime prevention already associates blackness with inherent criminality, why would we fight to make our faces more legible to a system designed to police us? … It is not social progress to make black people equally visible to software that will inevitably be further weaponized against us.”*

Reported by Sigal Samuel once again in the VOX article “**Facebook will finally ask permission before using facial recognition on you**,” it is becoming increasingly alarming to the average consumer, as well as to legal experts and AI ethicists, that large companies are using Deep Learning algorithms in ways that jeopardize the privacy of their users.

In some cases, such as for Facebook, these companies are being forced to allow customers to “Opt-In” to these services rather than to implement them by default or to use the “Opt-Out” model.

While this is a step in the right direction, it requires compliance by the companies, much of which appears to be unenforceable legally and reliant on good faith.

**Research Plan:**

1. Compete in Monet-ify Kaggle tournament to practice CGAN workflow and research tools (paperspace, PyTorch).
2. Look into Steganography via CGAN.

**Limitations:**

I don’t yet have a rigorous understanding of networks. By virtue of this, my ability to describe networks using conventional notation is relatively low. However, I have completed extensive coursework in computer graphics and have completed projects using CNNs in the past that have had excellent results. I am going to be completing all of the available rigorous coursework on ML, DL, AI, and AI Ethics over the next two semesters, so I’ll become far more equipped rigorously.

If the scope of this project is too large, I’d be very interested in narrowing it to something more bite-size.