Face/Off

Physical Adversarial Attacks for Facial Recognition Systems

Ronald Lencevičius

Department of Computer Science
Department of Mathematics, CPP
Faculty Advisors: Dr. Tingting Chen and Dr. Hao Ji, CPP
NSF REU Site: Big Data Security and Privacy 2019

Abstract

Facial recognition is becoming a widely used form of authentication and surveillance. Like most machine learning driven tools, face recognizers are prone to adversarial attacks. A successful attacker could be misidentified as someone else to bypass authentication or fool surveillance which poses a security risk. Sharif et al. demonstrated such a physical attack using perturbed eyeglasses which would be effective at fooling a live facial recognition system [1]. The aim of our work is to improve on physical adversarial attacks by creating a face mask or physical stickers that could perform well from different poses.

Background

The deep face recognition pipeline consists of 5 main parts: face detection, face alignment, face processing, face feature extraction, and face matching [2]. Face detection essentially recognizes the presence of a human face in an image. The face is then aligned by cropping down to the detected face while also mapping facial landmarks. This is then fed into the deep face recognizer which processes the aligned faces either for feature extraction or matching. The output for matching is an embedding of deep features that is then compared to a database of preexisting embeddings to see which one a particular face is close to.

In our case, an adversarial attack for facial recognition would be an input embedding that was perturbed along this pipeline in order to maximize the distance from the input's true identity while minimizing the distance between a target embedding that does not match the identity of the input.

Related Work

Face Detection [3]

MacDonald demonstrated that by generating histograms of oriented gradients (HOGs) using simulated annealing he could create a physical mask that could fool a popular HOG based detector (figure 1). This is the initial inspiration for our current work as our main goal is to generate a similar type of physical attack.

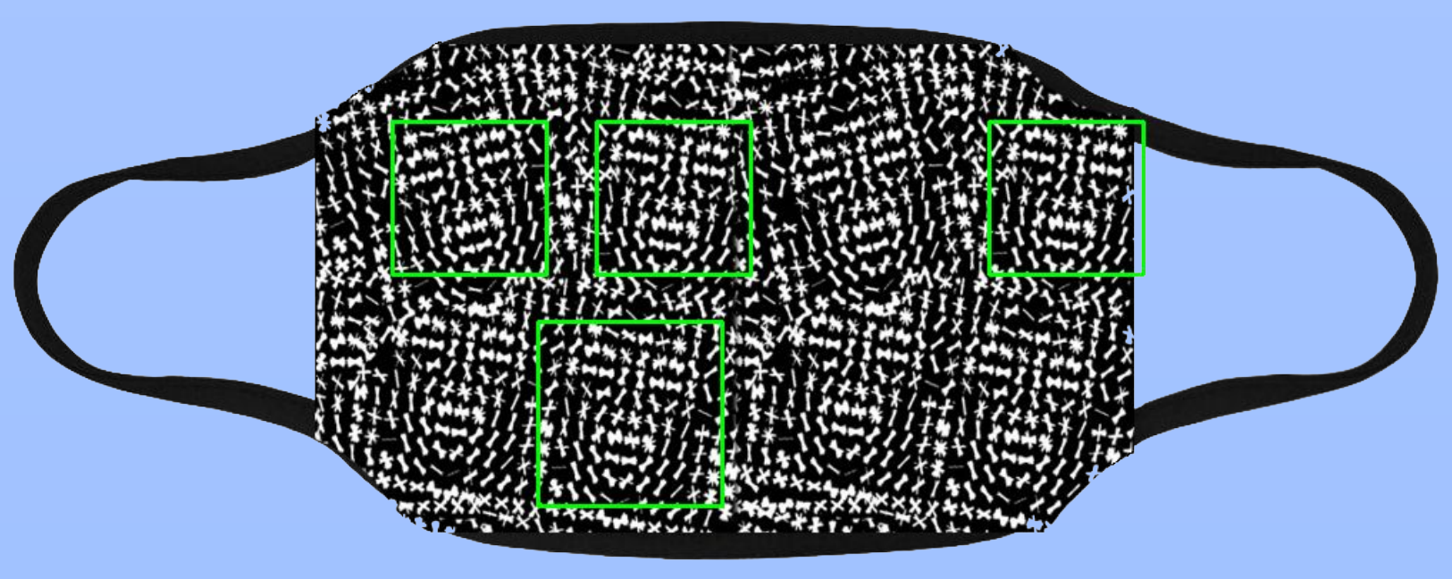


Figure 1: Facemask with HOG faces printed on it [3]. dlib's face detector was able to detect 5 different faces on this mask.

Physical adversarial attacks [4]

Thys et al. generated physical adversarial patches that could be worn by a person that could fool an object detector from recognizing the attacker as a person (figure 2). They used a modified euclidean distance loss function which took into consideration the ability to print such an example as well as consistent color transitions.

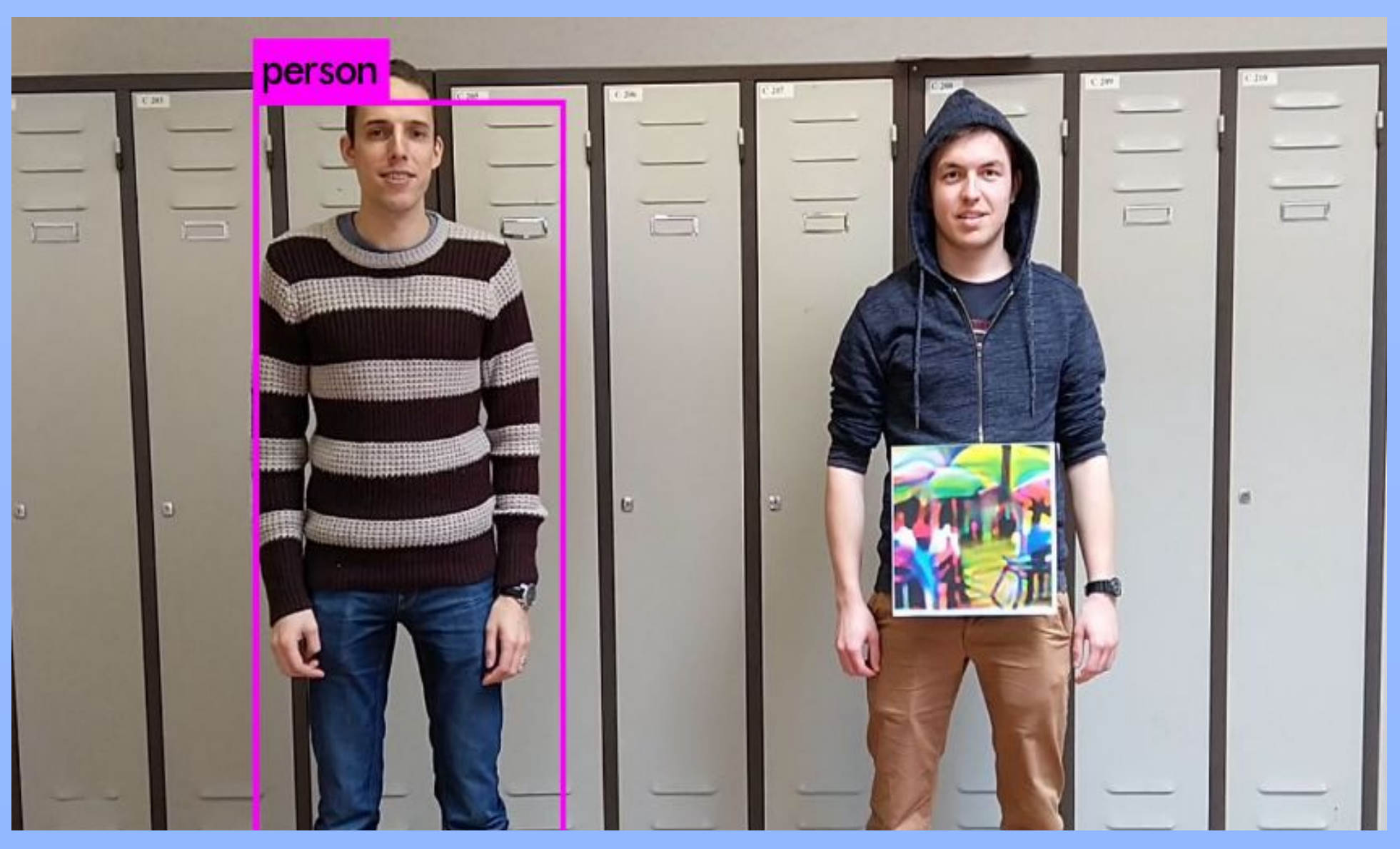
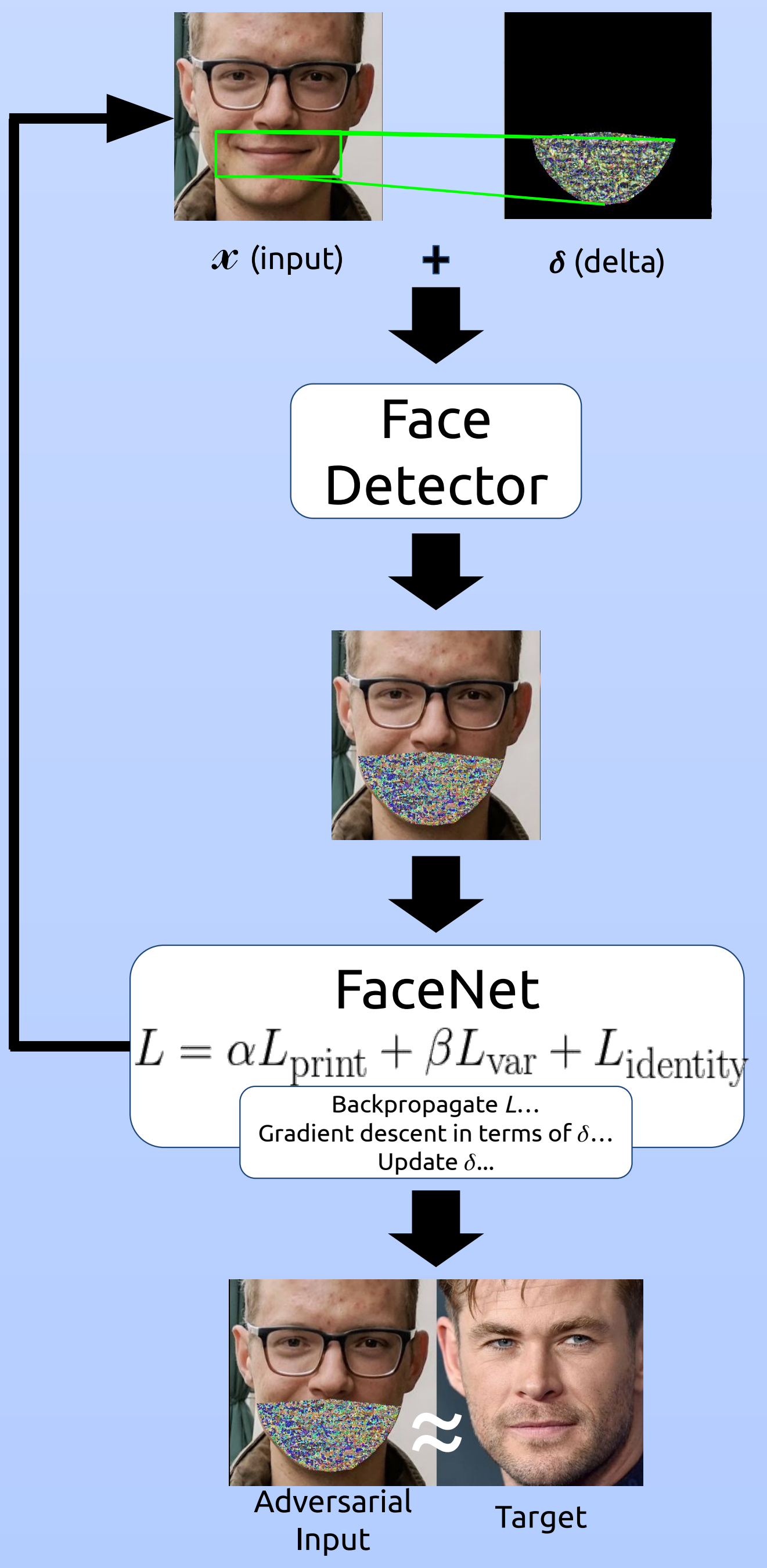


Figure 2: An example of an adversarial patch fooling a person detector [4].

Procedure



FaceNet

$$L = \alpha L_{\text{print}} + \beta L_{\text{var}} + L_{\text{identity}}$$

Backpropagate L…
Gradient descent in terms of δ…
Update δ…

Our current face recognition pipeline consists of an input image being fed into a face detector/aligner using dlib's histogram of oriented gradients detector for processing. This input is then normalized and fed into a face recognizer called FaceNet which outputs a face embedding.

What we want to accomplish is to perturb the input image around the mouth area, to generate enough of a disturbance for the face recognizer to generate an embedding that maximizes the distance between the input's true identity and minimizes the distance between the target.

We also want this perturbation to be viable for real world application that could be potentially printable as a mask or as a sticker. This would be accomplished by considering additional variables in our loss function as demonstrated by Thys et al [4].

Results

(Table 1)	True	Target
True	0.711	1.099
Adver.	1.113	0.164

(Table 2)	Input	Target	Transfer
Experiment 1	1.00	1.00	0.60
Experiment 2	1.00	0.00	0.00

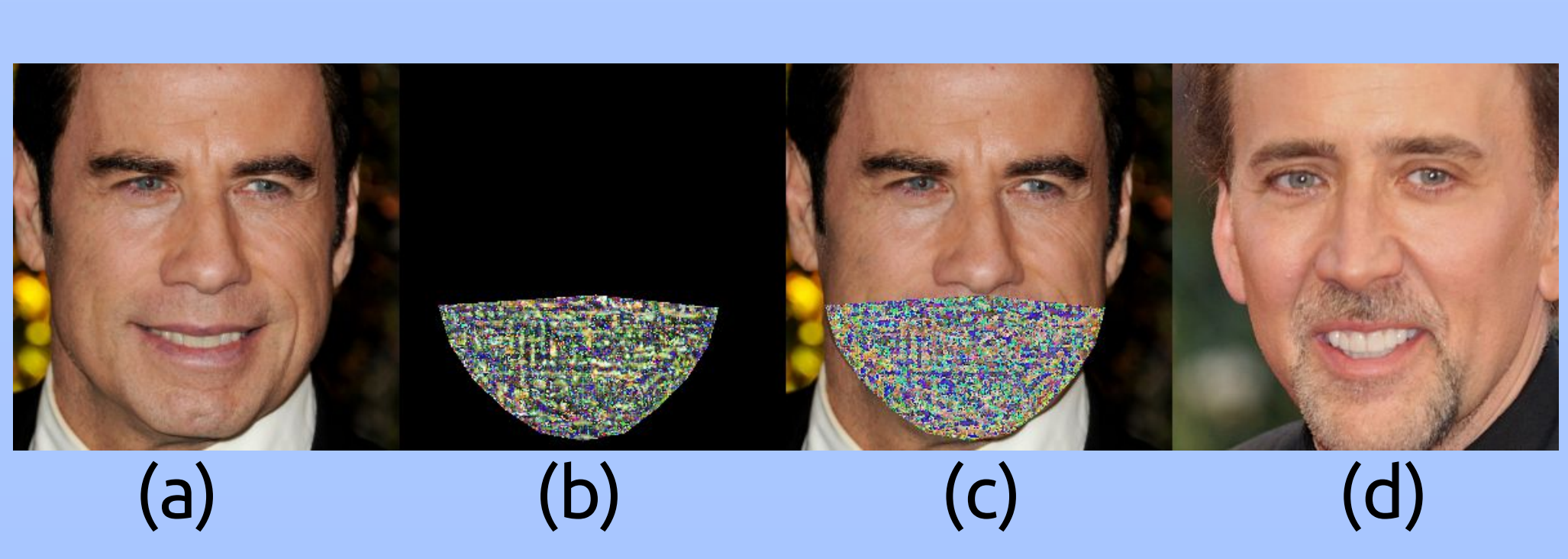


Table 1: In our current findings we were able to maximize the distance between the true input (a) and adversarial input (c) while **minimizing** the distance between the adversarial input (c) and the target identity (d) over 45 epochs of training

Table 2: Experiment 1 tested the performance of adversarial input for single input/single target as well as transferability for different target image. Experiment 2 tested the performance of adversarial input for the same input identity but different image with the same target identity. 5 identities were used for input/target with 2 images per ID

Future Work

- Work on improving the loss function to create physical adversarial attacks that are viable for printing
- Increase performance of the attack for different face poses
- Create a defense to increase robustness of the face recognition pipeline

References

[1] Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.
[2] Wang, Mei, and Weihong Deng. "Deep face recognition: A survey." arXiv preprint arXiv:1804.06655 (2018).
[3] MacDonald, Bruce. "Fooling Facial Detection with Fashion." Towards Data Science, Towards Data Science, 4 June 2019, towardsdatascience.com/fooling-facial-detection-with-fashion-d668ed919eb.
[4] Thys, Simen, et al. "Fooling automated surveillance cameras: adversarial patches to attack person detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.