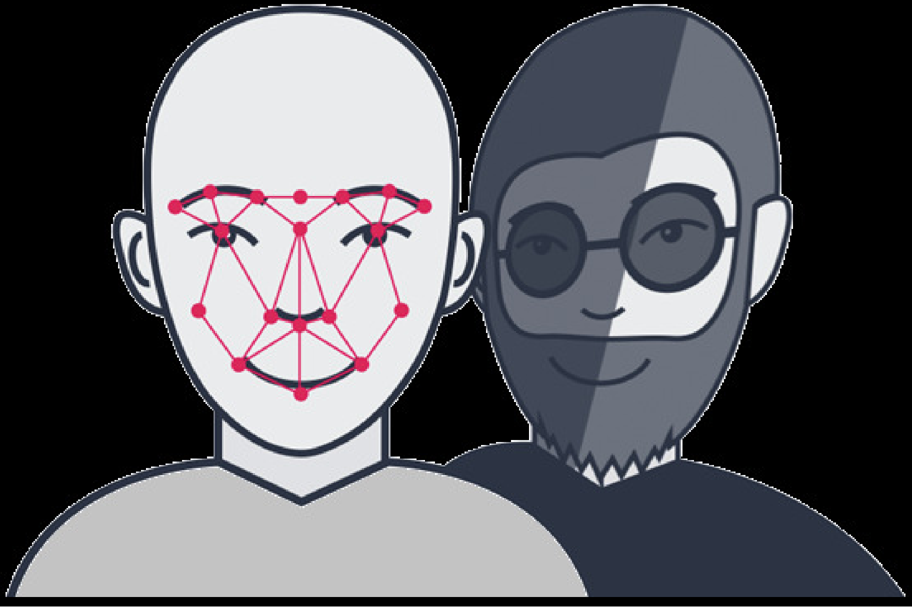
Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition



Keshav Yerra (2670843)

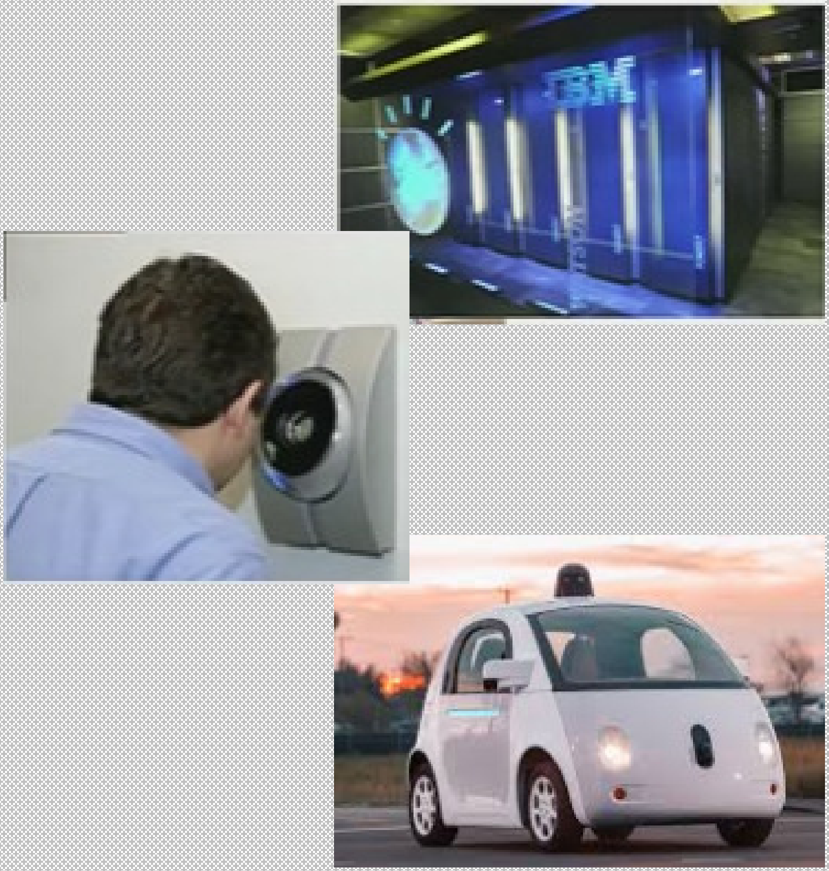
Monish Prasad (2671587)

Machine Learning is Everywhere

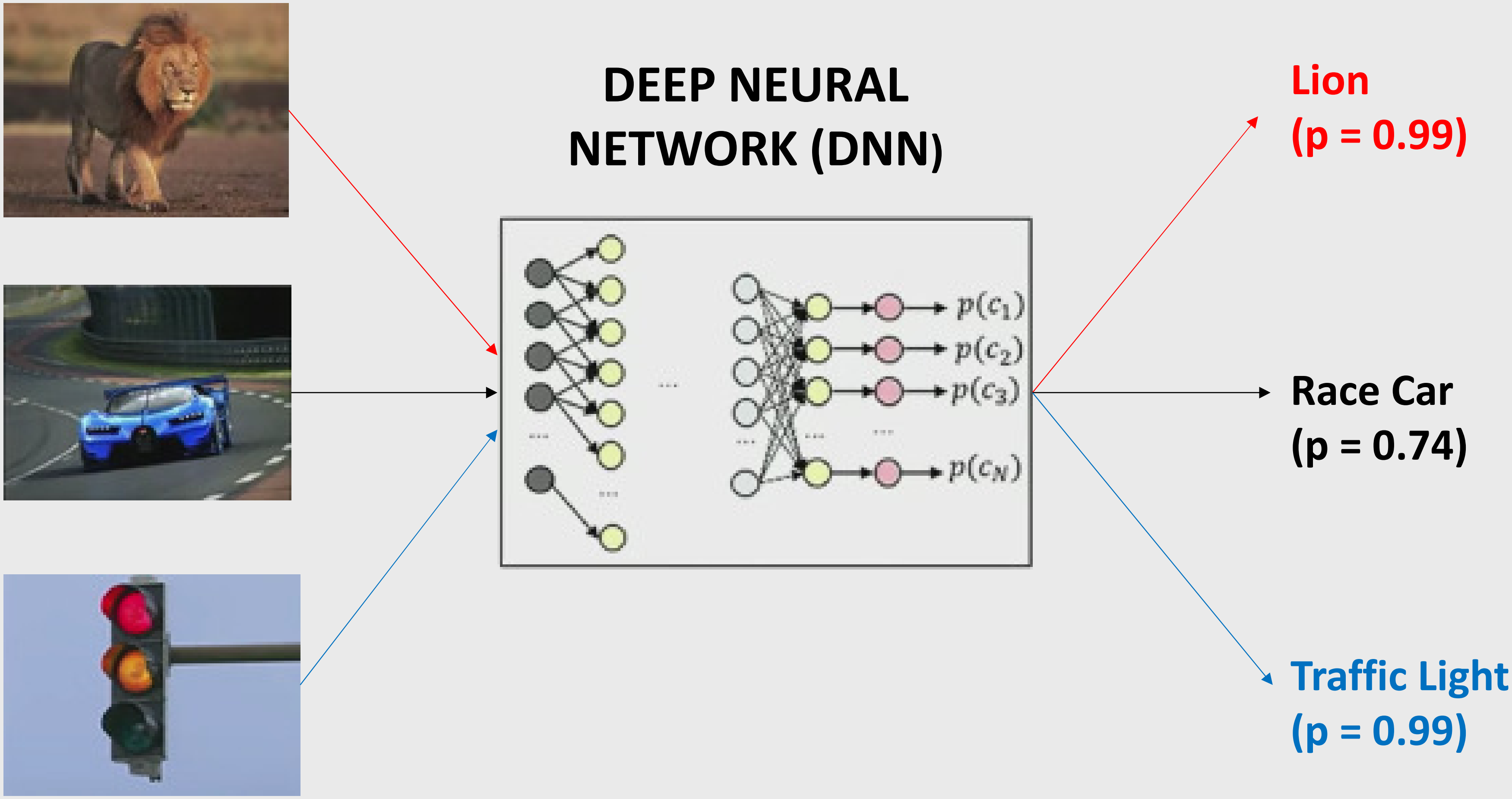
- Cancer Diagnosis

- Surveillance and access-control

- Self Driving Cars



What do we see here?



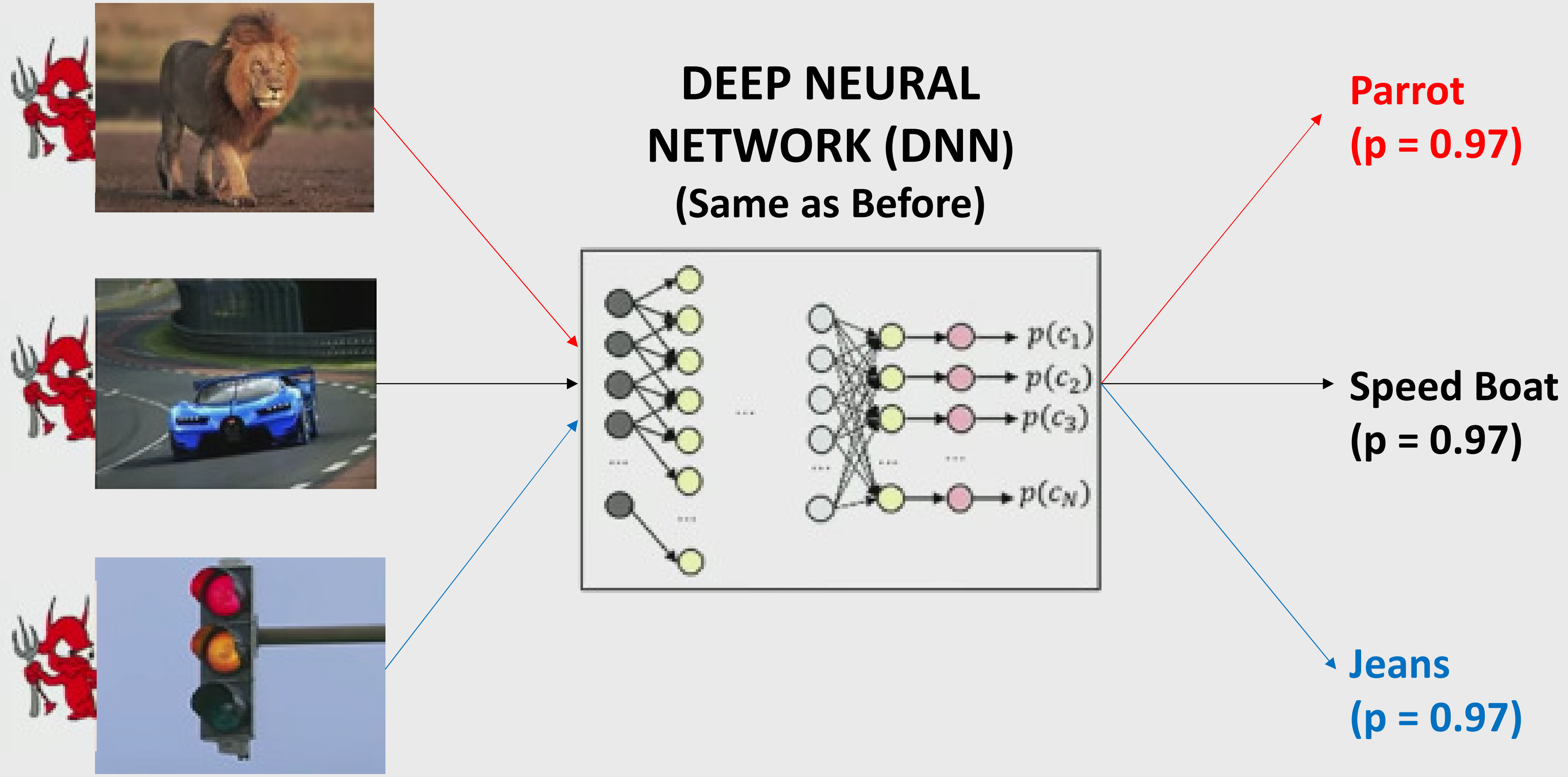
DEEP NEURAL NETWORK (DNN)

Lion
(p = 0.99)

Race Car
(p = 0.74)

Traffic Light
(p = 0.99)

What do we see here now?



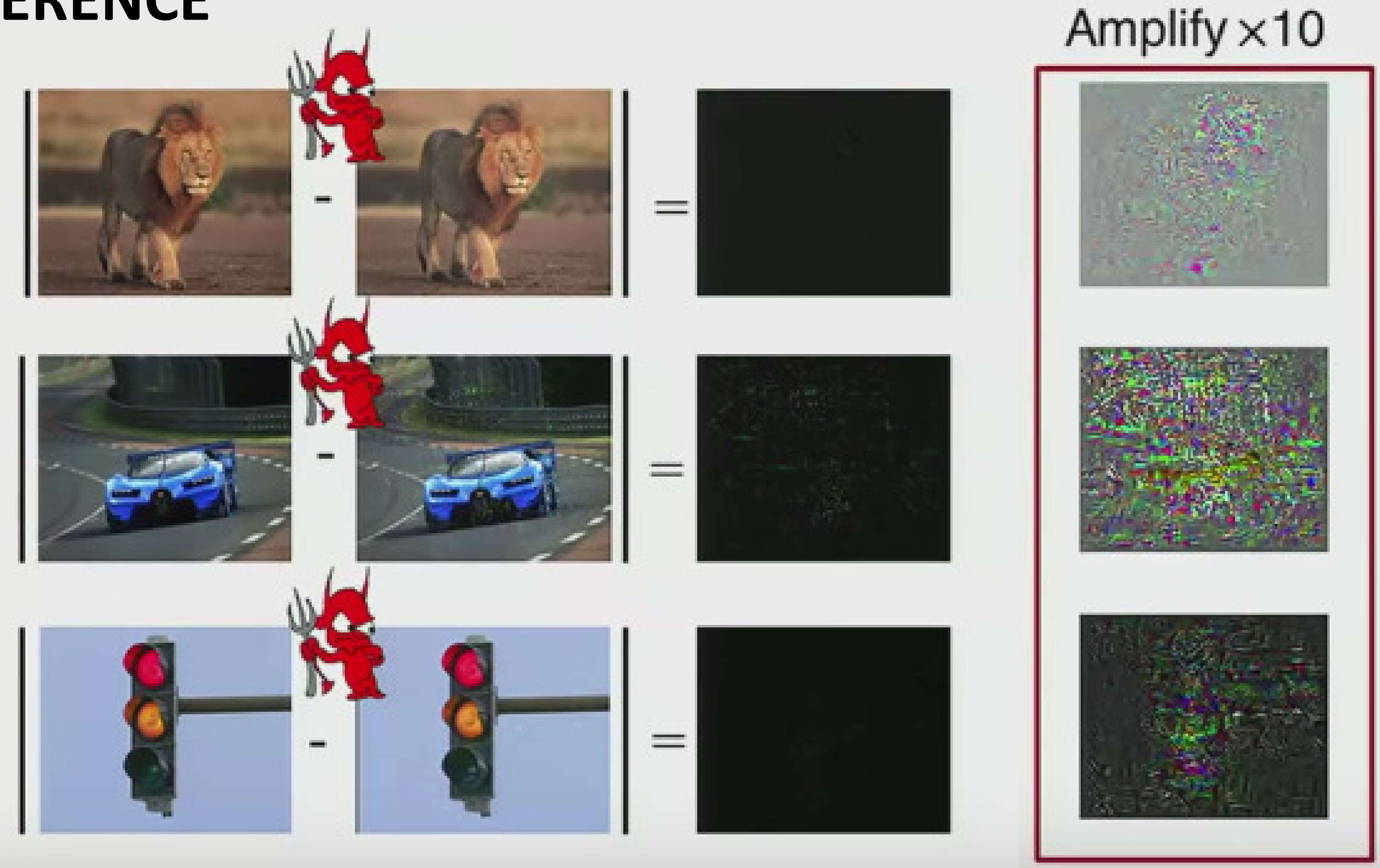
DEEP NEURAL NETWORK (DNN)
(Same as Before)

Parrot
(p = 0.97)

Speed Boat
(p = 0.97)

Jeans
(p = 0.97)

DIFFERENCE



Main Aim

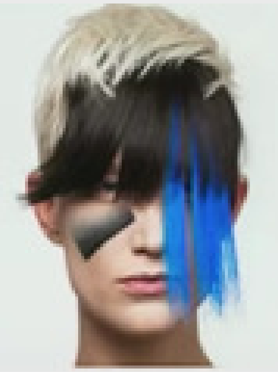
- The main aim is to find out whether an attacker will be able to successfully impersonate a victim in the real world by using some accessories to change his appearance.

Physical Realizability:

- Attacker can only change his own appearance
- Robust to changes in different imaging conditions

Inconspicuousness:

- Do not raise too much of suspicion
- Avoid physical appearances like



What are the adversary’s capabilities?

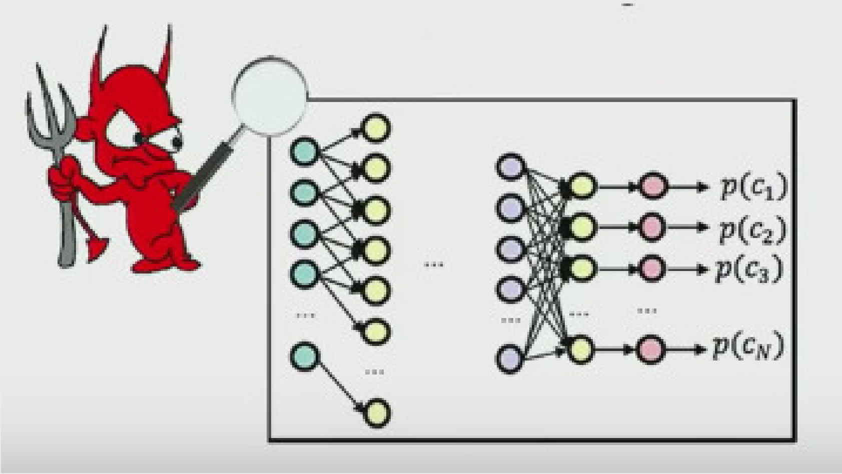
There are two types of settings:

- White-Box setting
- Black-Box setting

To generate attacks the attacker needs to know how changing input changes the output

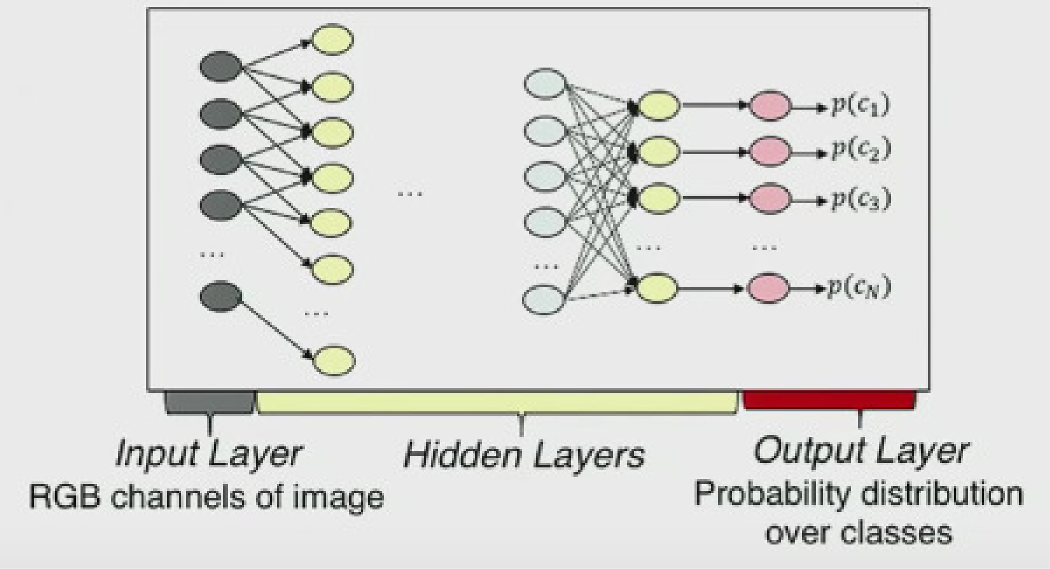
⬇

White-Box setting



What is a Deep Neural Network (DNN)?

- The basic idea is to stimulate how the brain cells work
- The basic building block is Neuron (A simple Computational Unit)



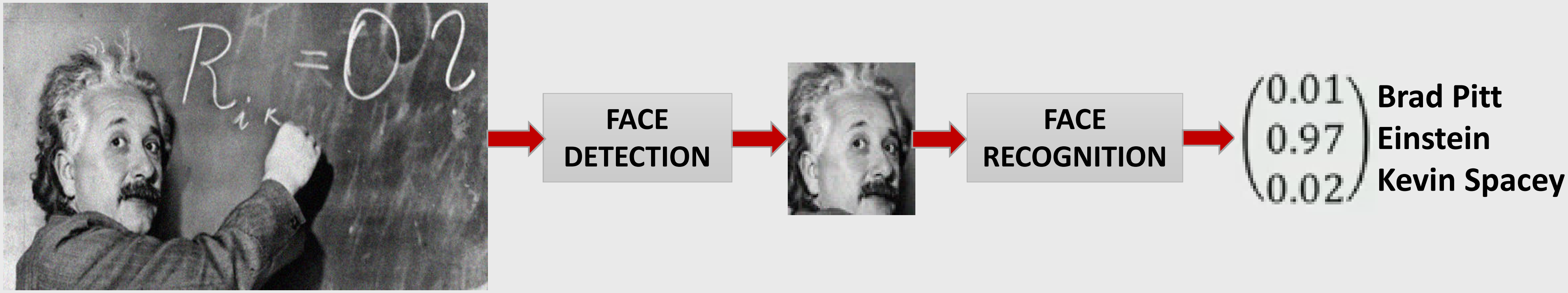
How to mislead this DNN?

- Given a DNN and input, we have to find a minimal change that causes a specific misclassification

Face Recognition

- Applications : surveillance, access control, ..
- Detection and Recognition are usually pipelined:

1. Detect the face
2. Recognize the person



Albert Einstein

FACE DETECTION

FACE RECOGNITION

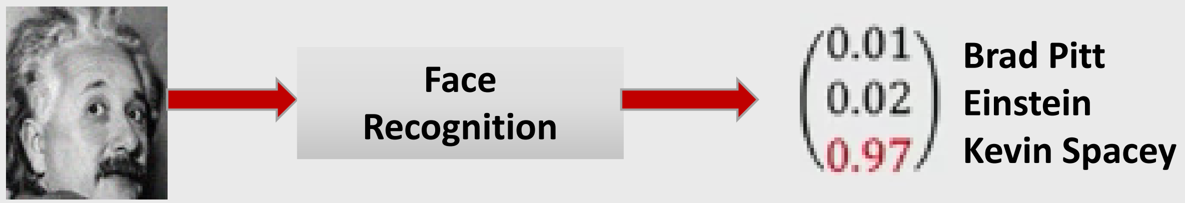
$$\begin{pmatrix}0.01\\0.97\\0.02\end{pmatrix}$$

Brad Pitt
Einstein
Kevin Spacey

Face Recognition Attacks

- Impersonation
- Dodging

Impersonation



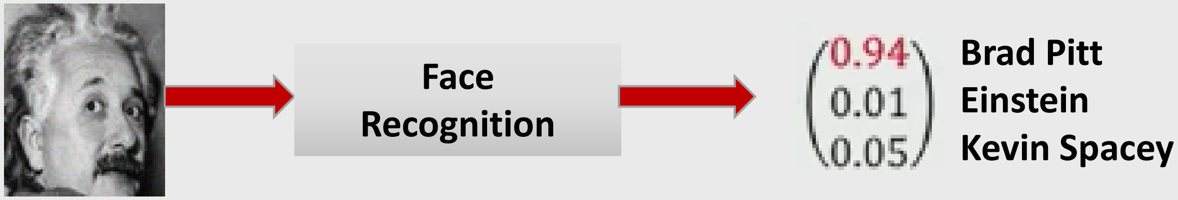
$$\left(\begin{array}{c}0.01\\0.02\\0.97\end{array}\right)\begin{array}{l}\textbf{Brad Pitt}\\\textbf{Einstein}\\\textbf{Kevin Spacey}\end{array}$$

Einstein

- Targeting a specific subject
- To access a specific resources or cause blame to be laid on the victim

Face Recognition Attacks

Dodging



Einstein

Face Recognition

$$\begin{array}{cl} \begin{pmatrix} 0.94 \\ 0.01 \\ 0.05 \end{pmatrix} & \begin{array}{l} \textbf{Brad Pitt} \\ \textbf{Einstein} \\ \textbf{Kevin Spacey} \end{array} \end{array}$$

- Being recognized incorrectly
- To Hide your Identity (or) If you don’t care who the victim is

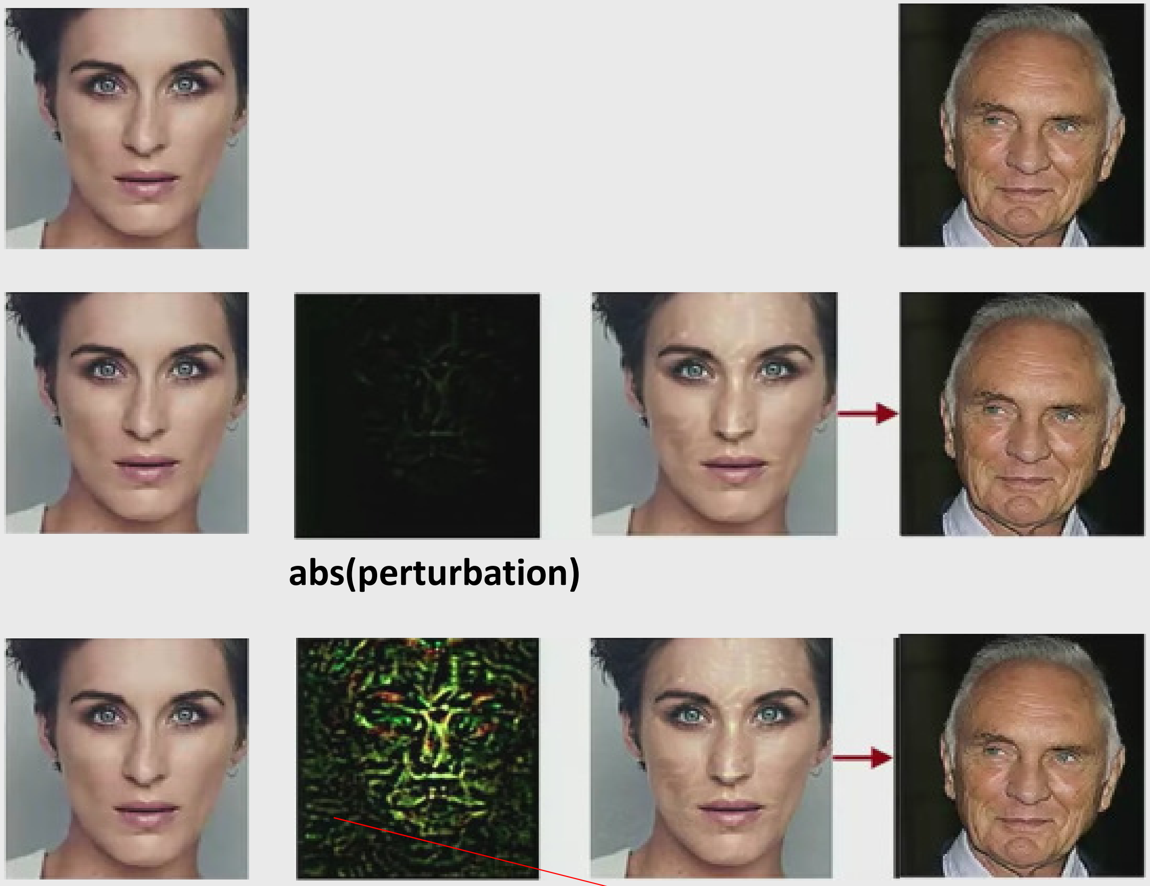
Deep Face Recognition

Here the DNN is built based on Parkhi et al. from [BMVC ’15]:

- The DNN built is trained to recognize 2622 celebrities
- About 13233 face images collected in the wild which are uncontrolled images
- And it outperforms humans:

Accuracy of humans	**Accuracy of Parkhi et al.’s DNN**
97.53%	**98.95%**

Example of impersonation:



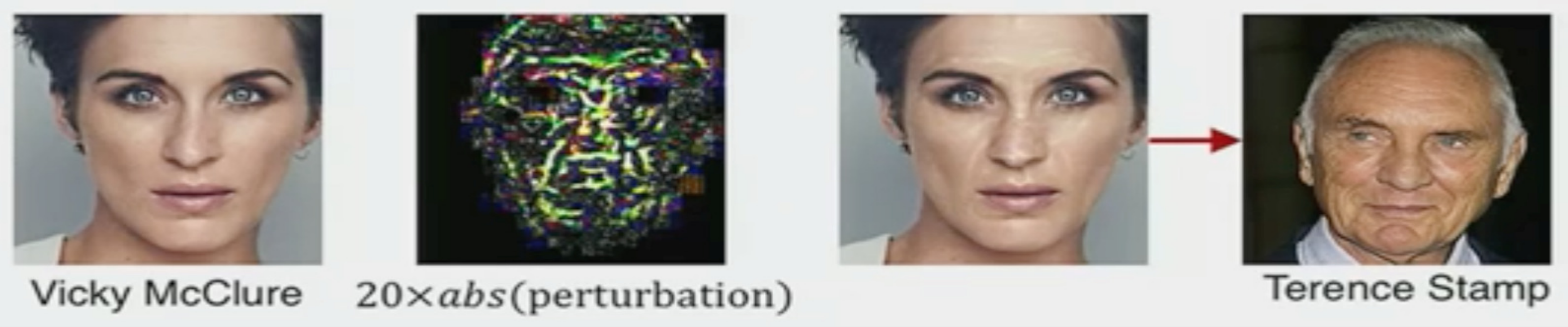
abs(perturbation)

10 x abs(perturbation)

The only problem for the attacker is controlling the background

PHASE #1: APPLY CHANGS TO THE FACE ONLY

- Image segmentation to find the face.
- Only change pixels that overlay the face.

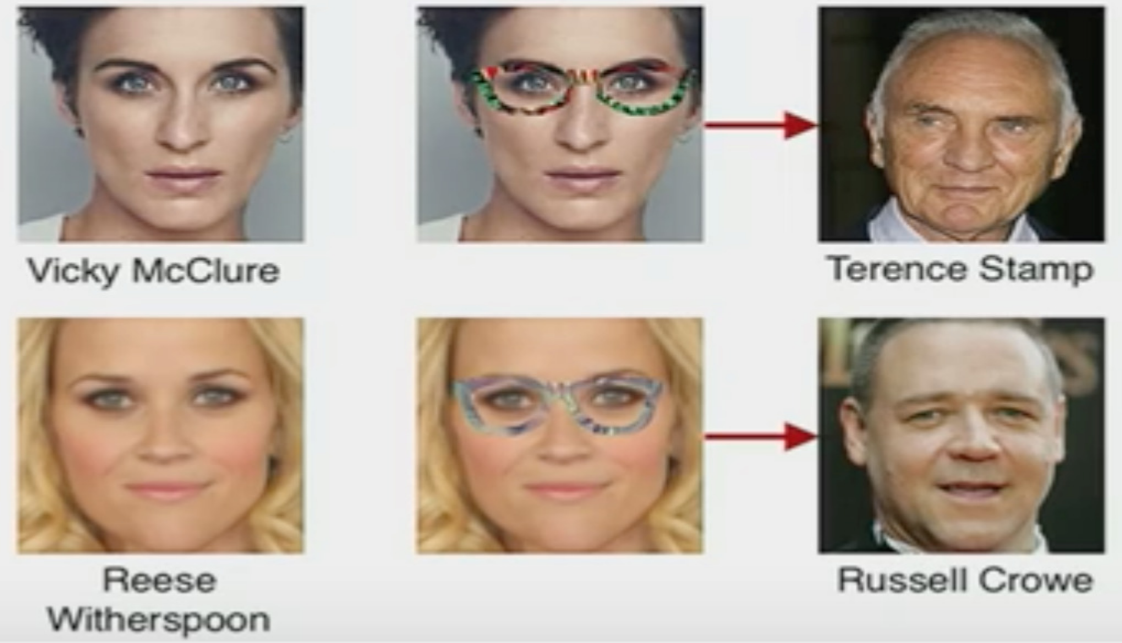


- Every impersonation attempt works.
- CAVEATS:

1. May be hard to realize the perturbations.
2. Perturbations are smaller than the camera’s sampling error.

PHASE #2: APPLY CHANGES TO THE EYEGLASSES

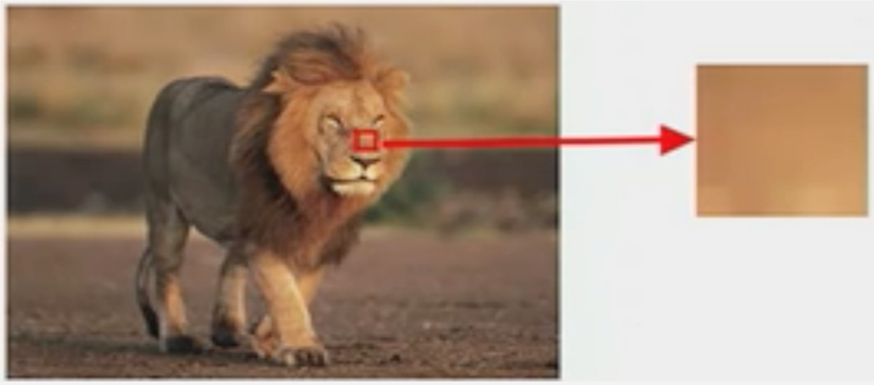
- Easier to realize(2D or 3D printing)
- Wearing eyeglasses isn’t associated with adversarial intent.



Impersonation Attempts success rate : 92%

PHASE #3: SMOOTH TRANSITIONS

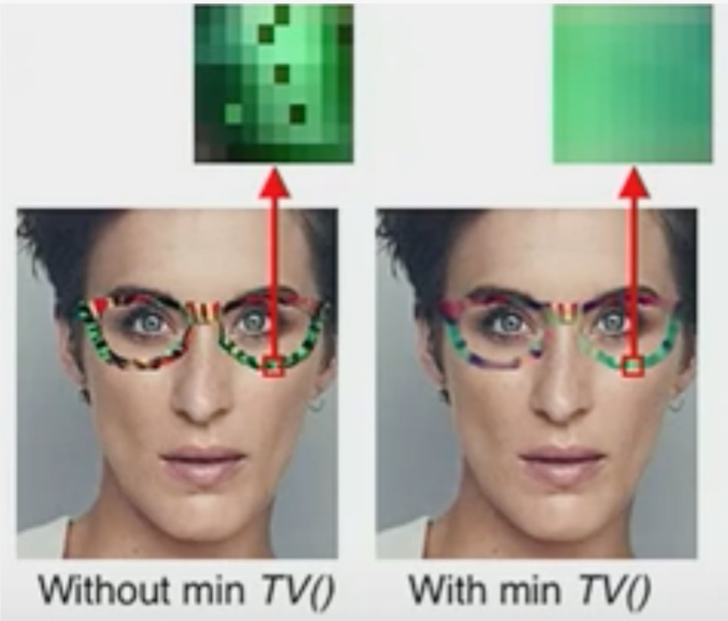
- Natural images tend to be smooth.



- We achieve this by minimizing the total variations:

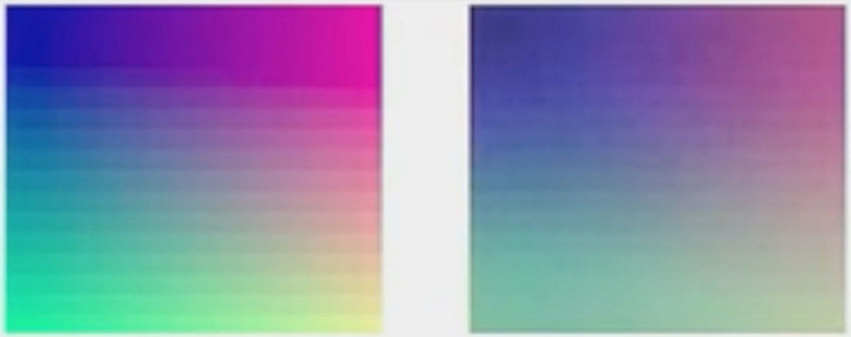
$$\mathrm{TV}(r) = \boxed{\sum_{i,j}\sqrt{\left(r_{i,j+1}-r_{i,j}\right)^2+\left(r_{i+1,j}-r_{i,j}\right)^2}}$$

Sum of differences of neighboring pixels



PHASE #4: PRINTABLE EYEGALSSES

- Challenge: Cannot print all the colors.
- Find the printable colors by printing color pallets.



Ideal Color Palette

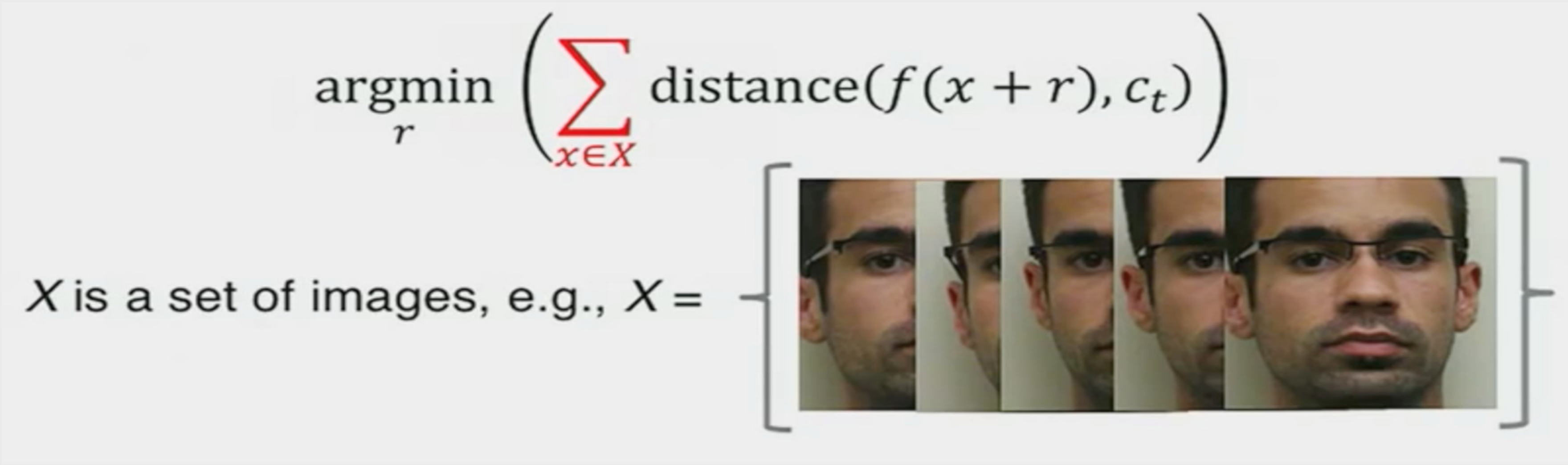
Printed Color Palette

- Define Non-Printability Score(NPS)

1. NPS is high if color is not printable
2. Generate printable eyeglasses by minimizing NPS

PHASE #5: ROBUST PERTURBATIONS

- Two samples of the same face are almost never the same.
- Attack should be generalized beyond one image.
- This is achieved by finding one attack accessory that leads any image in a set of images to be misclassified.



$$\underset{r}{\operatorname{argmin}} \left(\sum_{x \in X} \text{distance}(f(x+r), c_t) \right)$$

X is a set of images, e.g., $X =$

PUTTING IT ALL TOGETHER

- Physically realizable impersonation.

$$\underset{r}{\operatorname{argmin}} \underbrace{\boxed{\left(\sum_{x\in X} \text{distance}(f(x+r), c_t)\right)}}_{\substack{\text{misclassify as } c_t \\ \text{(set of images)}}} + \underbrace{\boxed{\kappa_1 \cdot \text{TV}(r)}}_{\text{smoothness}} + \underbrace{\boxed{\kappa_2 \cdot \text{NPS}(r)}}_{\text{printability}}$$

TESTING THE APPROACH:

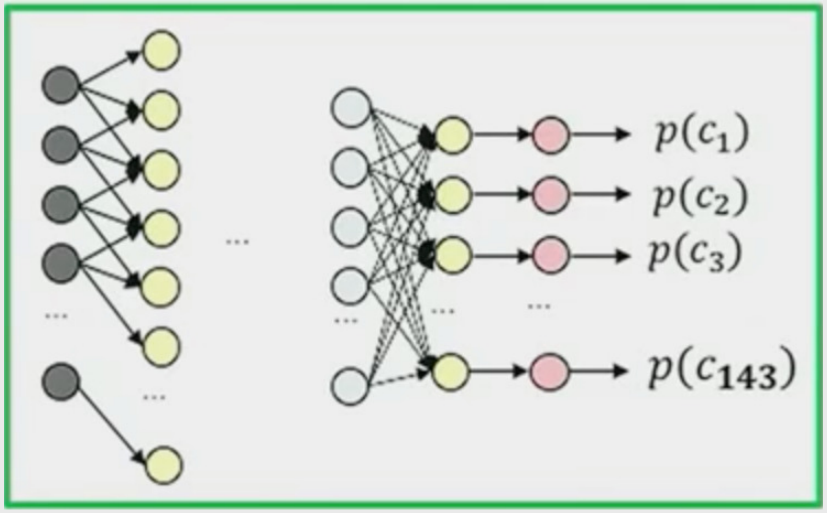
1. People to play role of the attacker.
2. Realize the eyeglasses.
3. DNN that recognizes the attackers.

DNN that Recognizes the Us

- It is hard to train DNN from scratch ➔ Use standard technique (transfer learning) to retrain DNN from Parkhi et al.'s
- New DNN recognizes 143 subjects:

(3 authors + 140 Celebs from PubFig dataset)

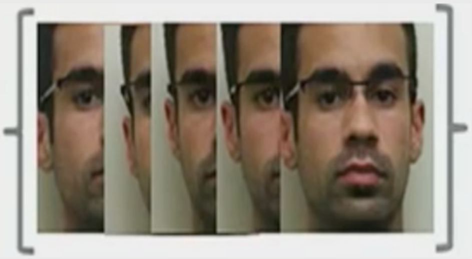
- Accuracy achieved: 96.75%

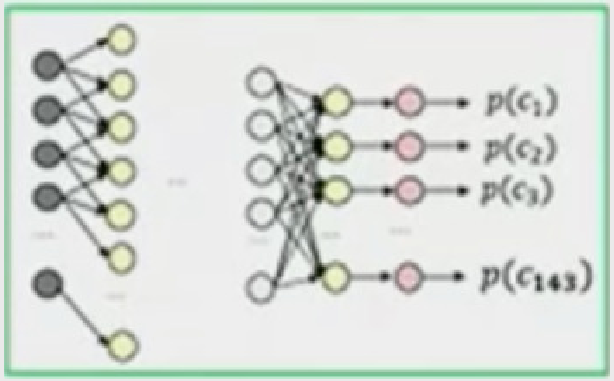


EXPERIMENT: REALIZED IMPERSONATIONS

PROCEDURE

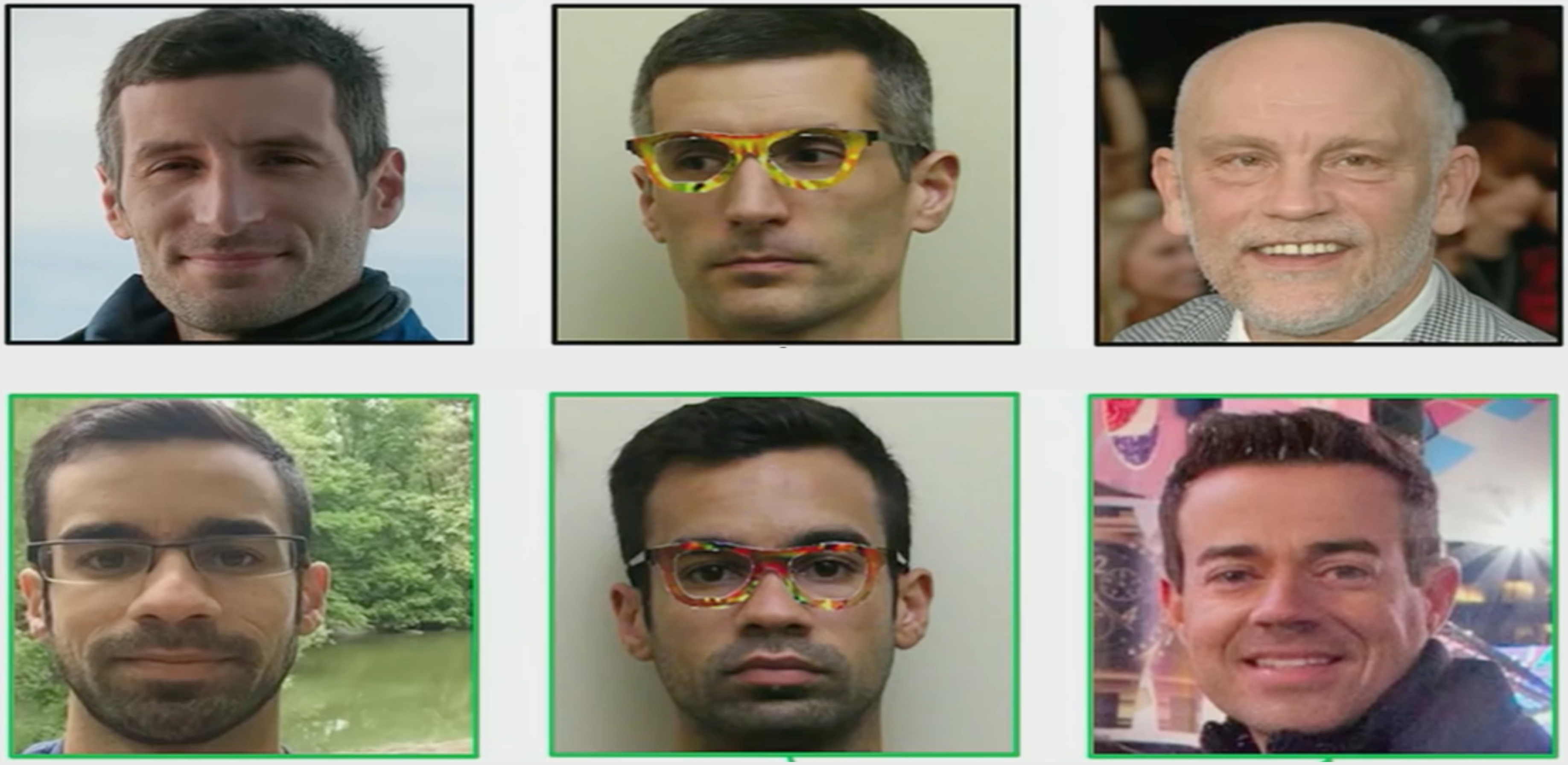
1. Collect images of attacker.
2. Chose random target.
3. Generate and print eyeglasses.
4. Collect 30 to 50 images of attacker wearing the eyeglasses.
5. Classify the collected images.





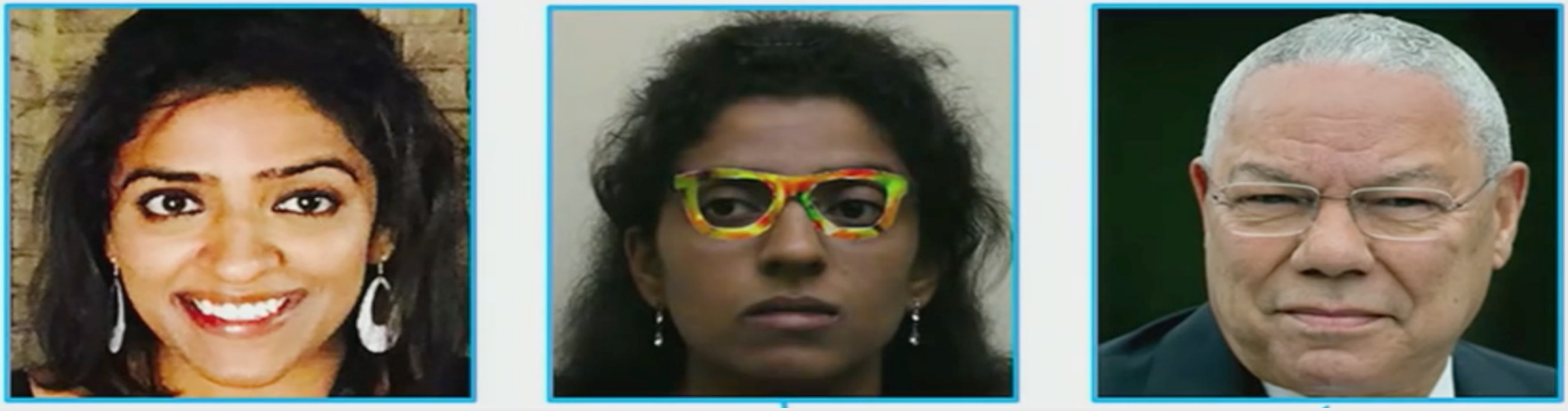
- Success Metric: Fraction of collected images misclassified as target.
- Limitation: Small set of variations in lighting

IMPERSONATION ATTACKS POSE A HUGE RISK:

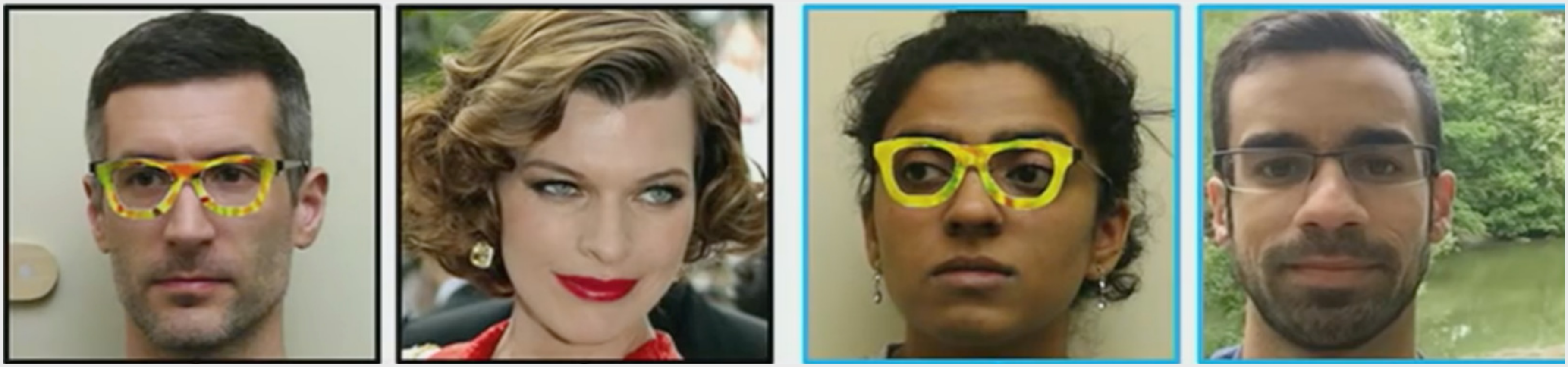


100% SUCCESS

- 16% SUCCESS:



- 88% SUCCESS:



CONCLUSION

1. Dodging and impersonation attacks can mislead state-of-the-art face recognition.

2. Attacks can be inconspicuous and physically realized.

3. Extensions to:

- Black-Box Models
- Invisibility against face detection.

THANK-YOU!!