Adversarial Machine Learning: Curiosity, Benefit, or Threat?

Lujo Bauer

Associate Professor
Electrical & Computer Engineering + Computer Science

Director
Cyber Autonomy Research Center

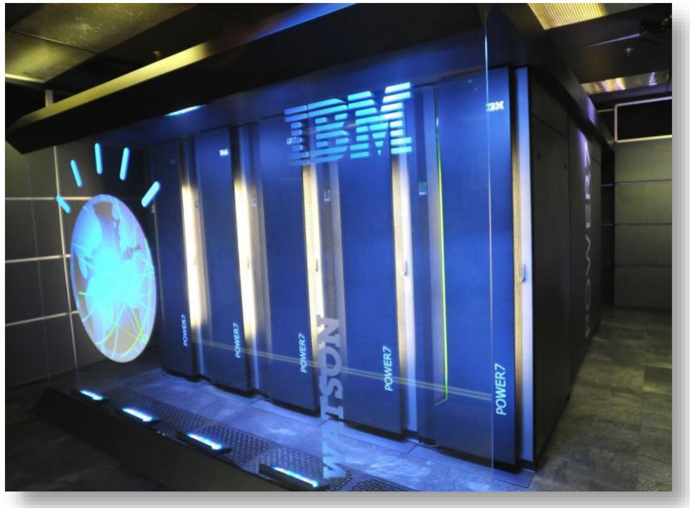
Collaborators:
Mahmood Sharif, Sruti Bhagavatula, Mike Reiter (UNC)

Carnegie Mellon University

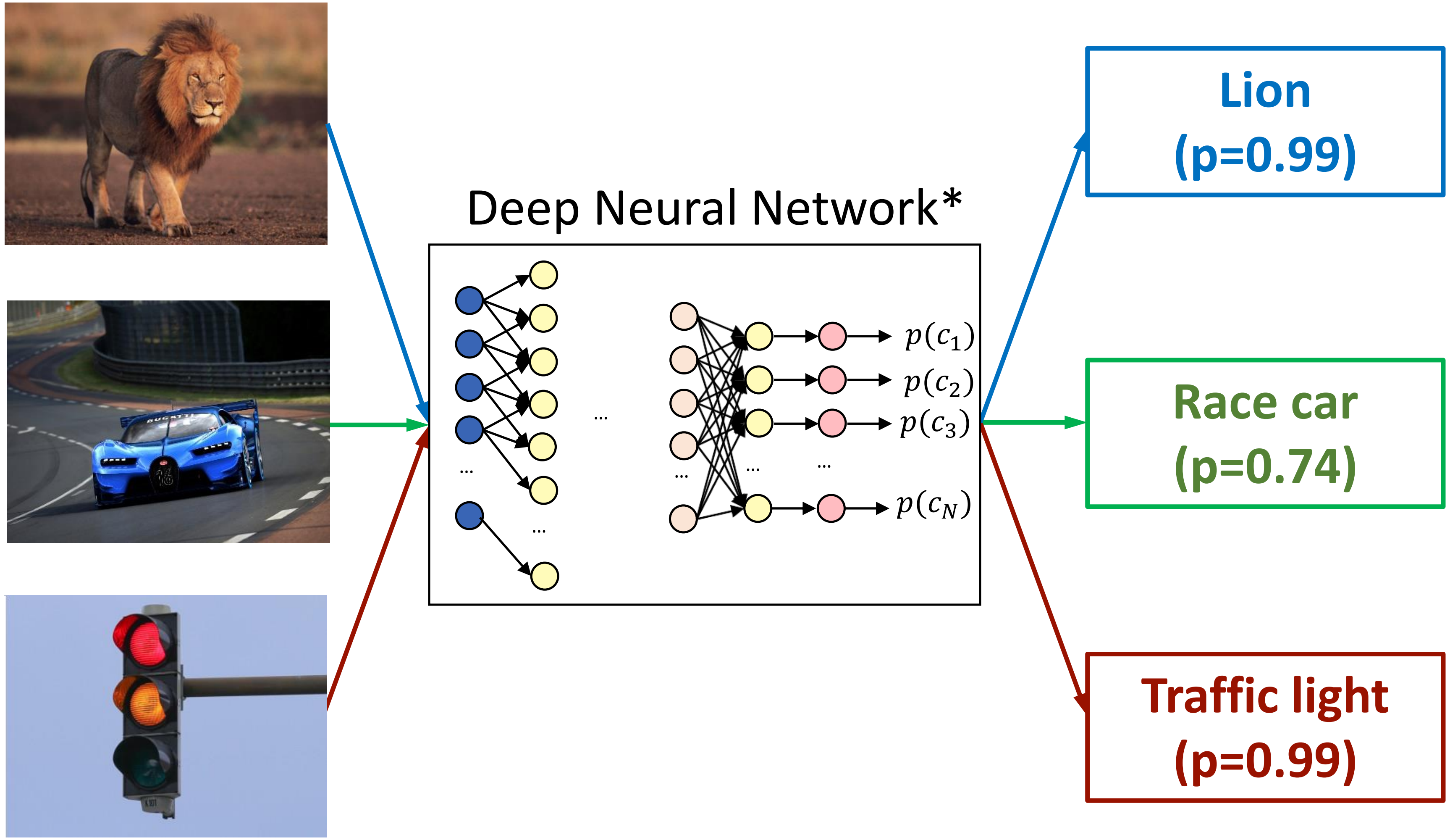
Carnegie Mellon University CyLab Security and Privacy Institute

Machine Learning Is Ubiquitous

- Cancer diagnosis
- Predicting weather
- Self-driving cars
- Surveillance and access-control

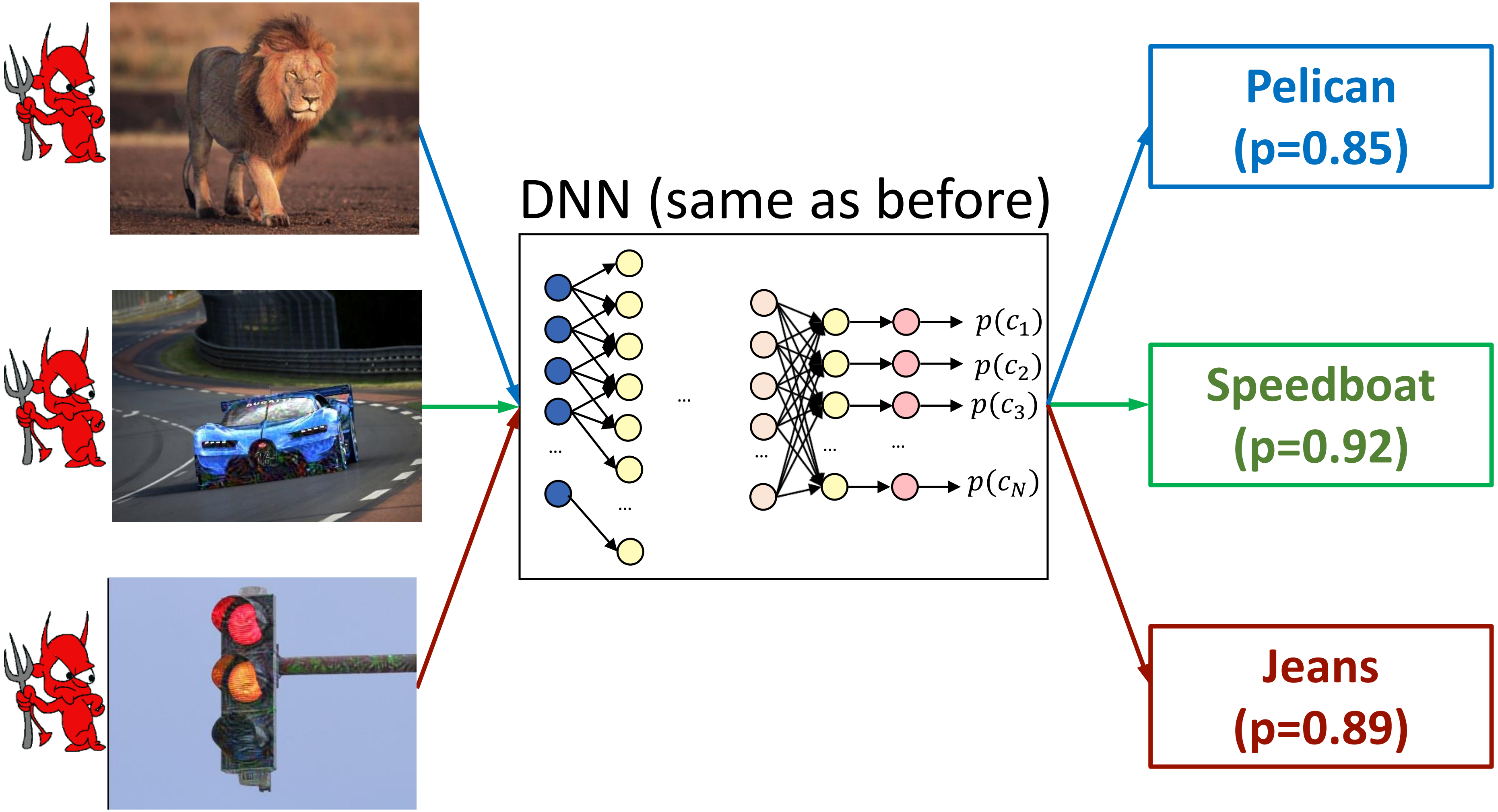


What Do You See?



*CNN-F, proposed by Chatfield et al., “Return of the Devil”, BMVC ‘14

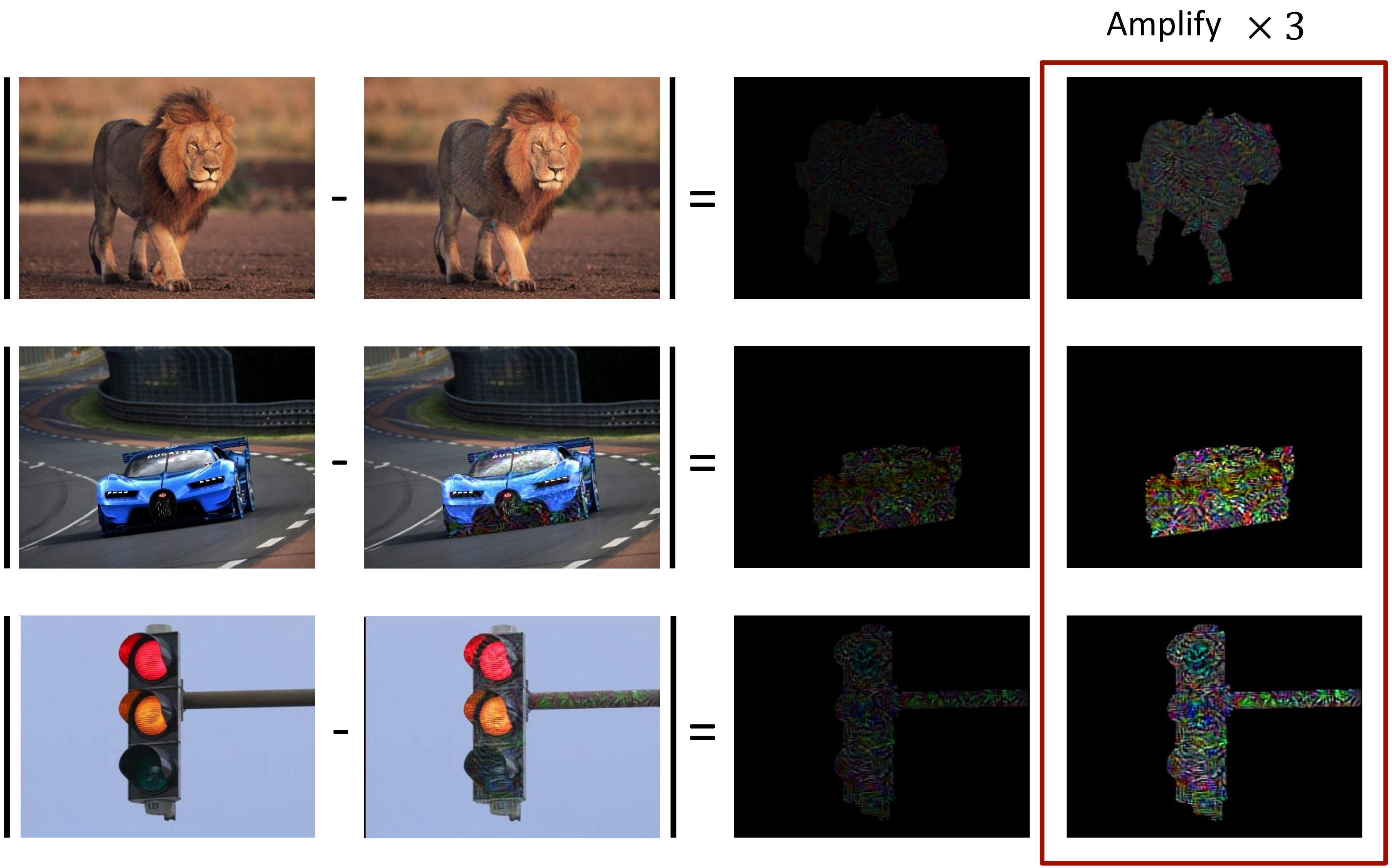
What Do You See Now?



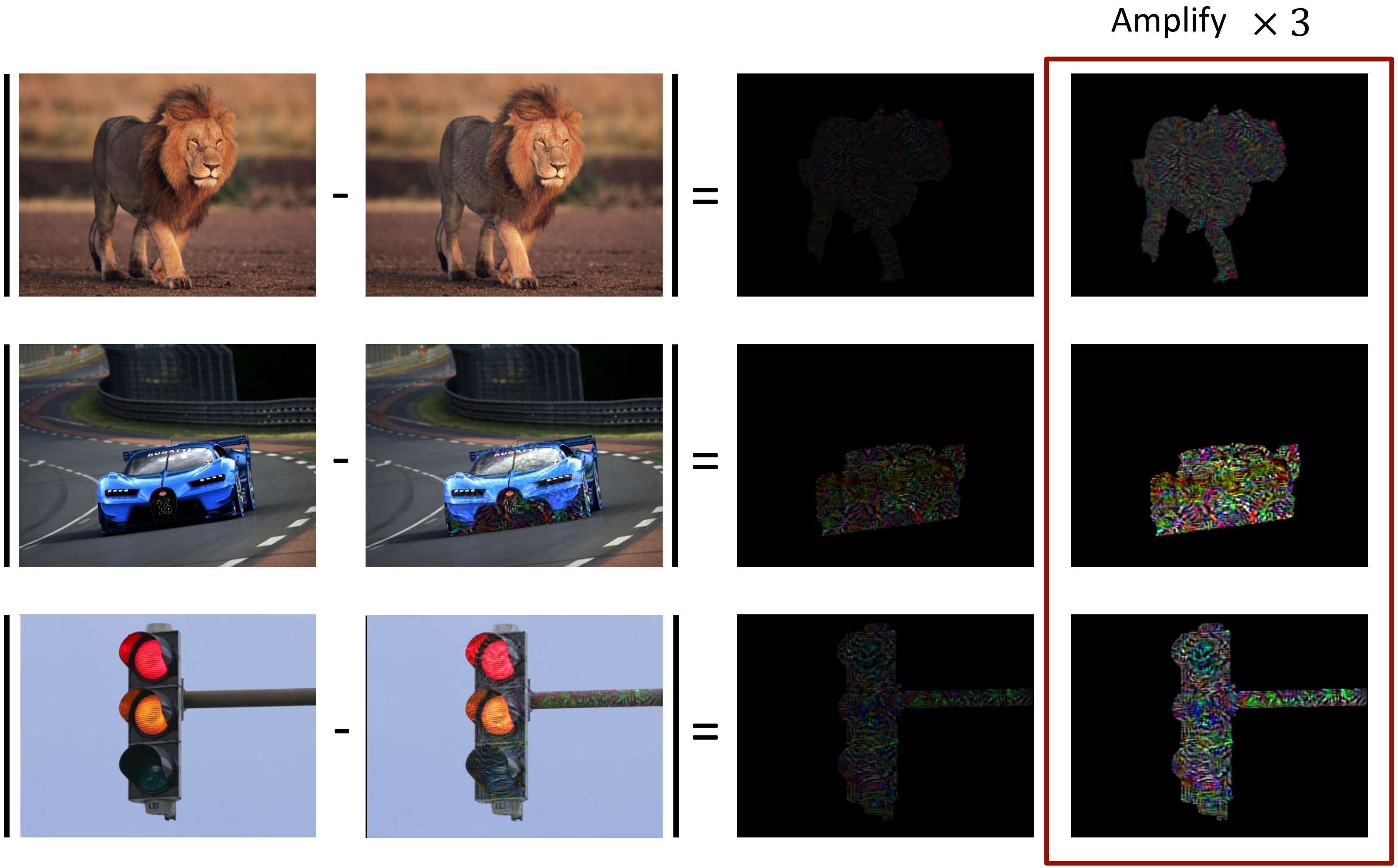
*The attacks generated following the method proposed by Szegedy et al.

Carnegie Mellon University CyLab Security and Privacy Institute

The Difference



Is This an Attack?



Can *an Attacker* Fool ML Classifiers?

[Sharif, Bhagavatula, Bauer, Reiter
CCS ’16, arXiv ’17, TOPS ’19]

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects, in a limited way

Can’t control camera position, lighting

Defender / beholder doesn’t notice attack
(to be measured by user study)

Attempt #1

0. Start with Szegedy et al.'s attack
1. Restrict modification to eyeglasses
2. Smooth pixel transitions
3. Restrict to printable colors
4. Add robustness to pose

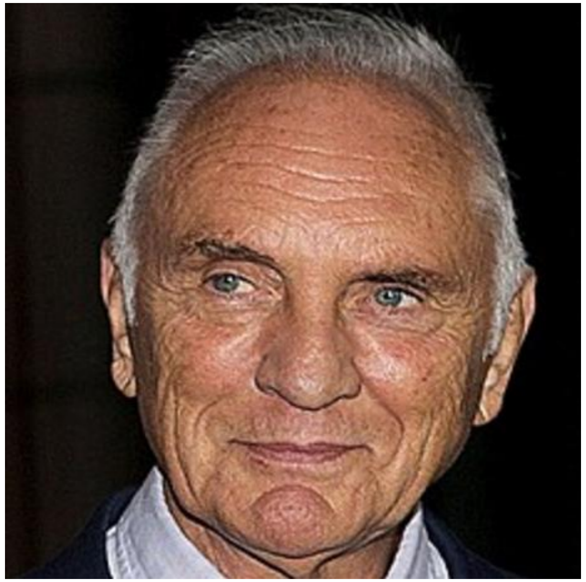
"Inconspicuousness"

Physical realizability

Carnegie Mellon University CyLab Security and Privacy Institute

Step #1: Apply Changes Just to Eyeglasses

Vicky McClure

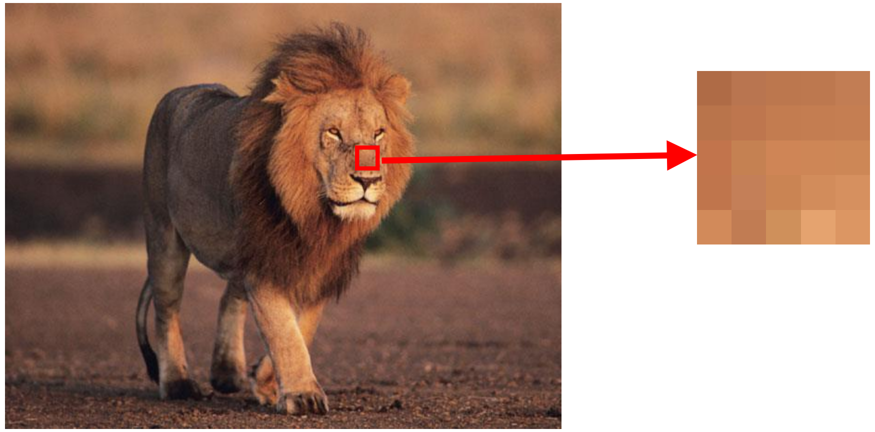


Terence Stamp

Carnegie Mellon University CyLab Security and Privacy Institute

Step #2: Smooth Pixel Transitions

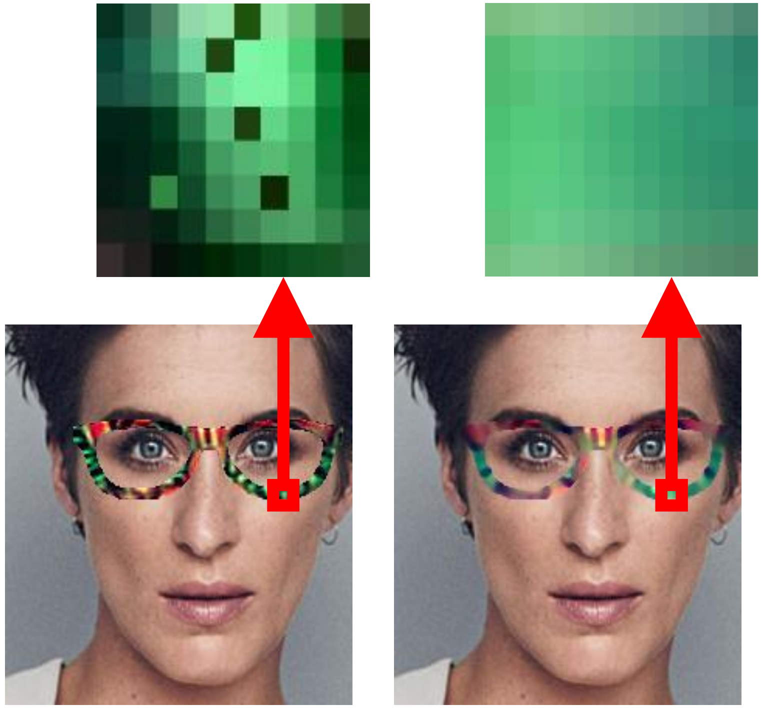
Natural images tend to be smooth:



We minimize 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



Without min *TV()* With min *TV()*

Step #3: Restrict to Printable Colors

- Challenge: Cannot print all colors
- Find printable colors by printing color palette

Ideal
color palette

Printed
color palette

- Define non-printability score (NPS):
  - high if colors are not printable; low otherwise
- Generate printable eyeglasses by minimizing NPS

Carnegie Mellon University
CyLab
Security and Privacy Institute

Step #4: Add Robustness to Pose

- Two samples of the same face are almost never the same ⇒ attack should generalize beyond one image
- Achieved by finding one eyeglasses that lead any image in a set 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* = 

Carnegie Mellon University CyLab Security and Privacy Institute

Putting All the Pieces Together

$$\underset{r}{\operatorname{argmin}} \boxed{\left(\sum_{x \in X} \operatorname{distance}(f(x+r), c_t)\right)} + \boxed{\kappa_1 \cdot \mathrm{TV}(r)} + \boxed{\kappa_2 \cdot \mathrm{NPS}(r)}$$

misclassify as c_t
(set of images)

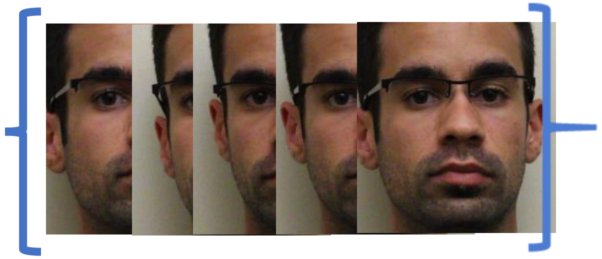
smoothness

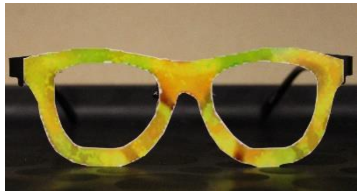
printability

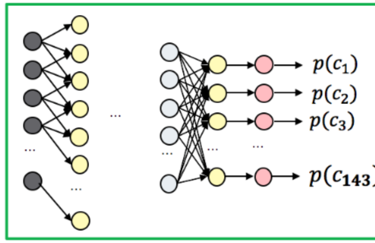
Time to Test!

Procedure:

0. Train face recognizer
1. Collect images of attacker
2. Choose random target
3. Generate and print eyeglasses
4. Collect images of attacker wearing eyeglasses
5. Classify collected images







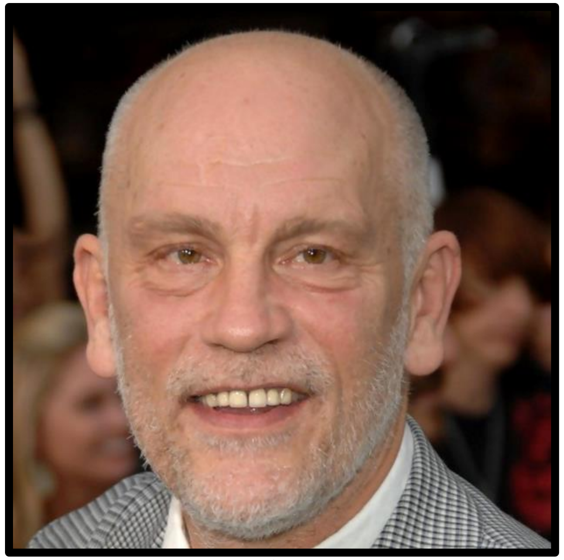
Success metric: fraction of images misclassified as target

Carnegie Mellon University
CyLab
Security and Privacy Institute

Physically Realized Impersonation Attacks Work

Lujo

John Malkovich



100% success

Physically Realized Impersonation Attacks Work

Mahmood

Carson Daly

100% success

Carnegie Mellon University
CyLab
Security and Privacy Institute

Can *an Attacker* Fool ML Classifiers? (Attempt #1)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects, in a limited way ✓

Can't control camera position, lighting ?

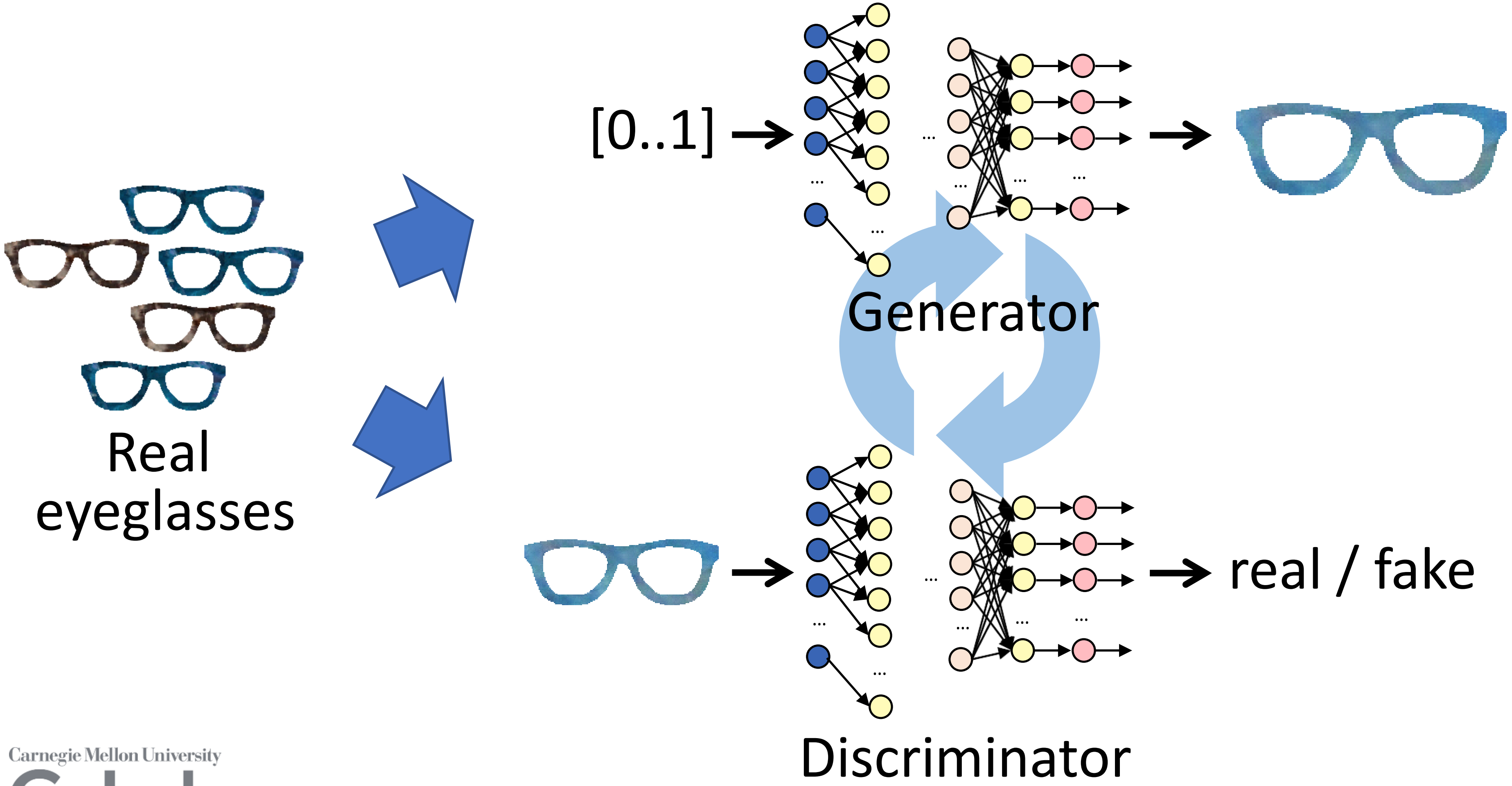
Defender / beholder doesn't notice attack ?
(to be measured by user study)

Attempt #2

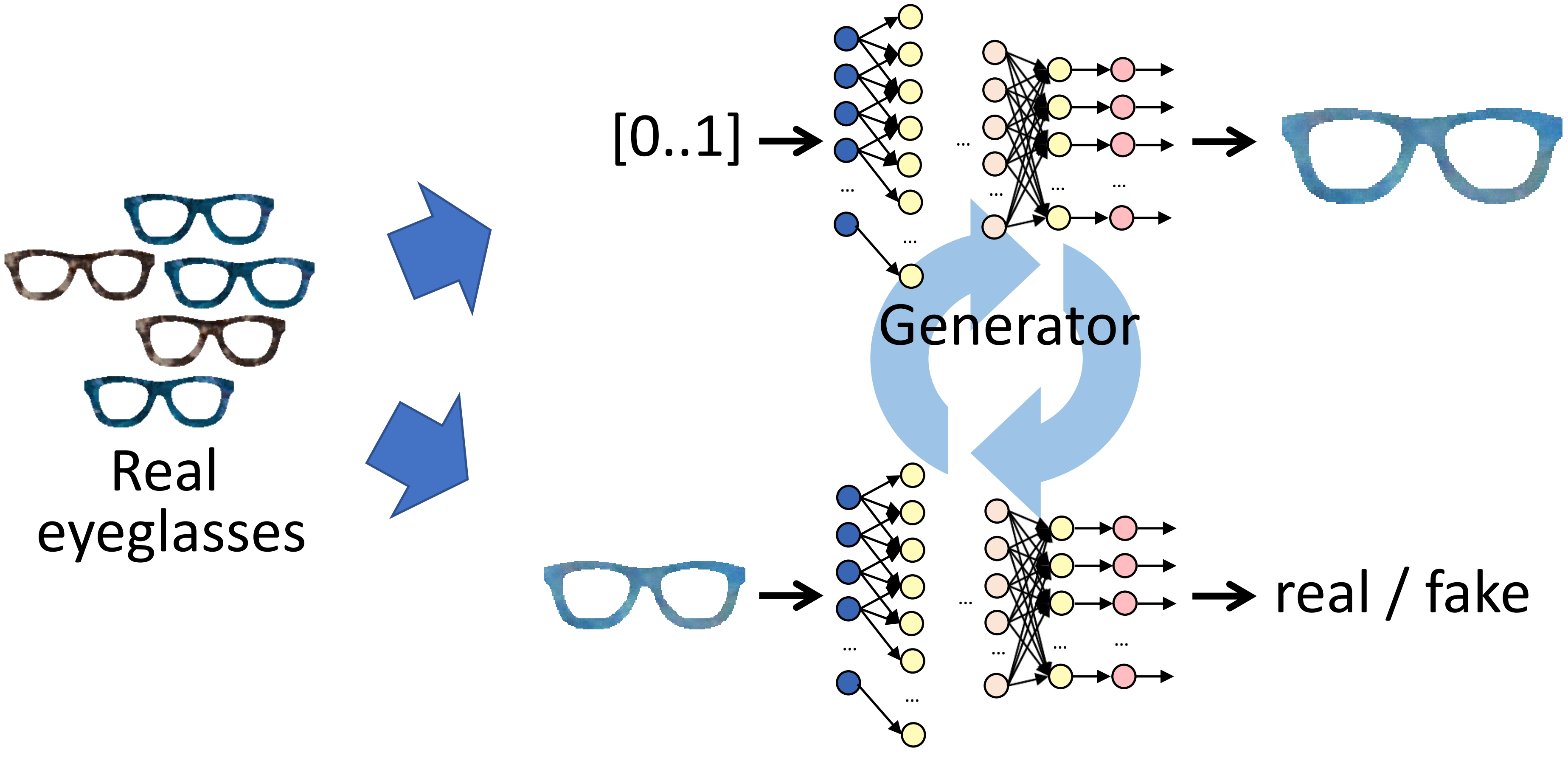
Goal: Capture hard-to-formalize constraints, i.e., "inconspicuousness"

Approach: Encode constraints using a neural network

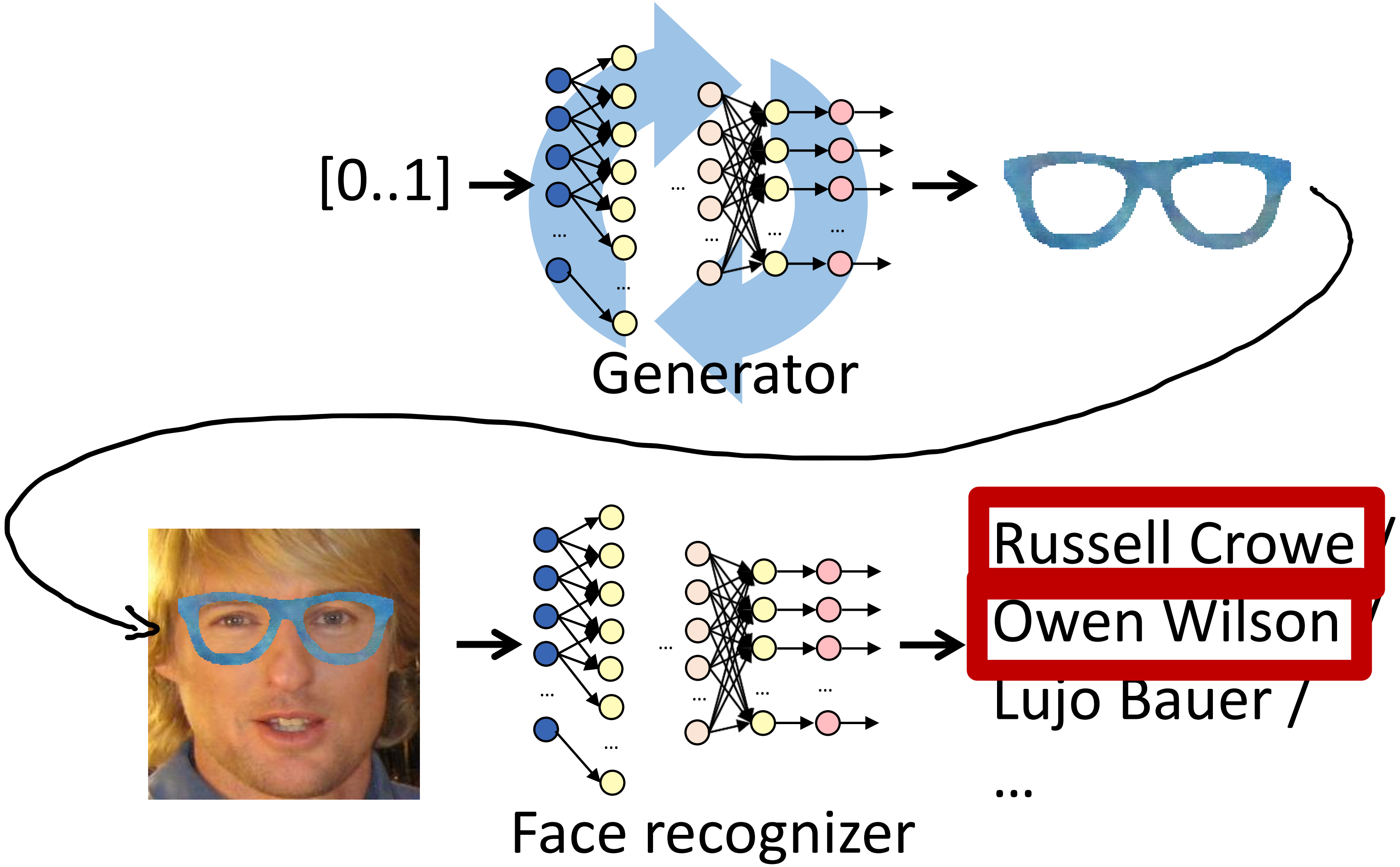
Step #1: Generate Realistic Eyeglasses

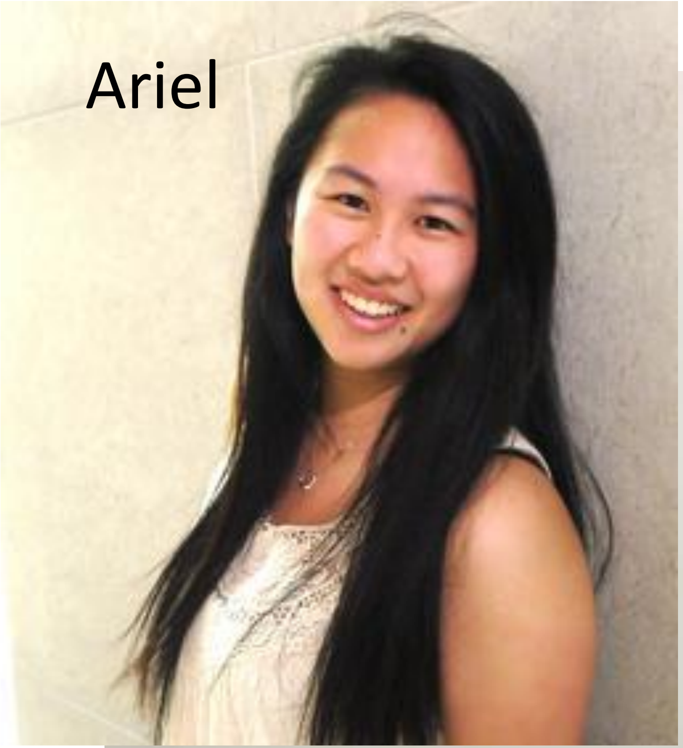


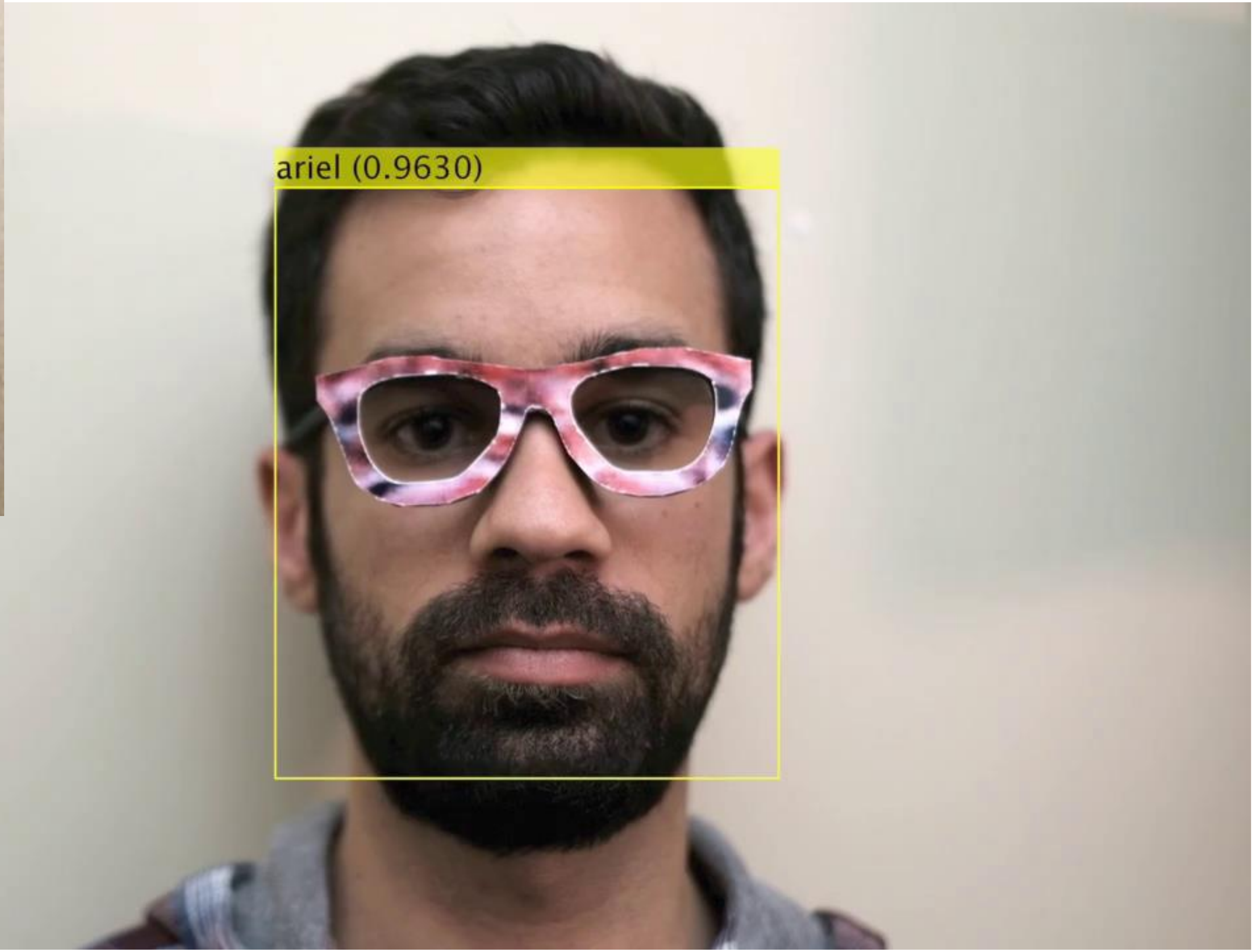
Step #2: Generate Realistic Eyeglasses
Adversarial



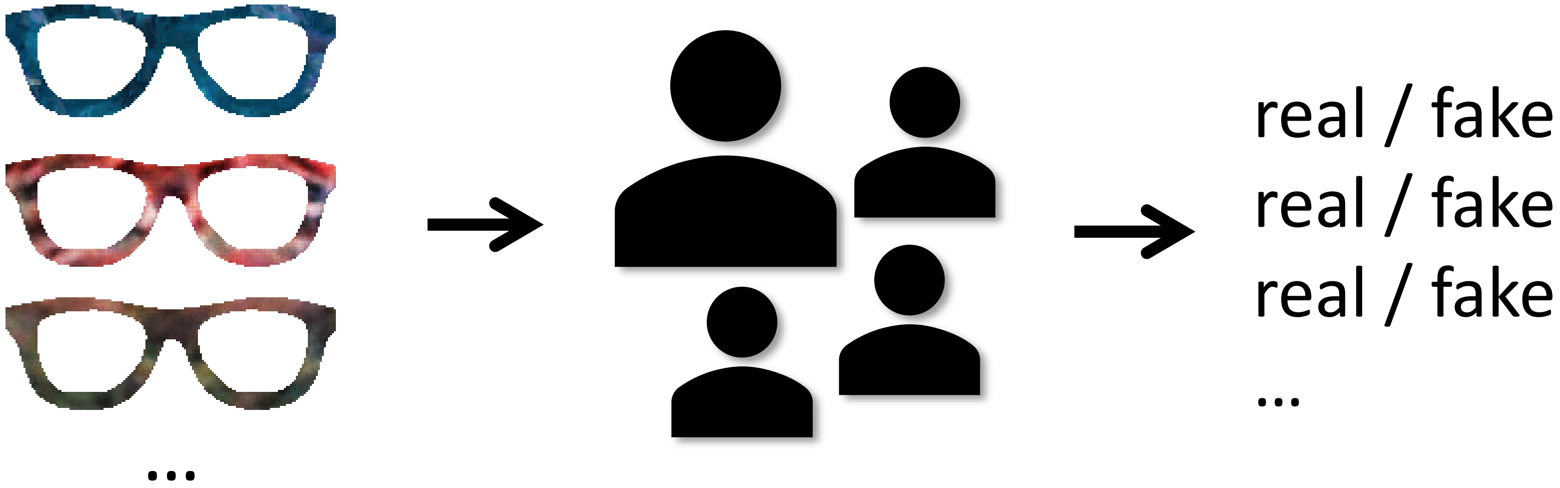
Step #2: Generate Realistic Eyeglasses
Adversarial



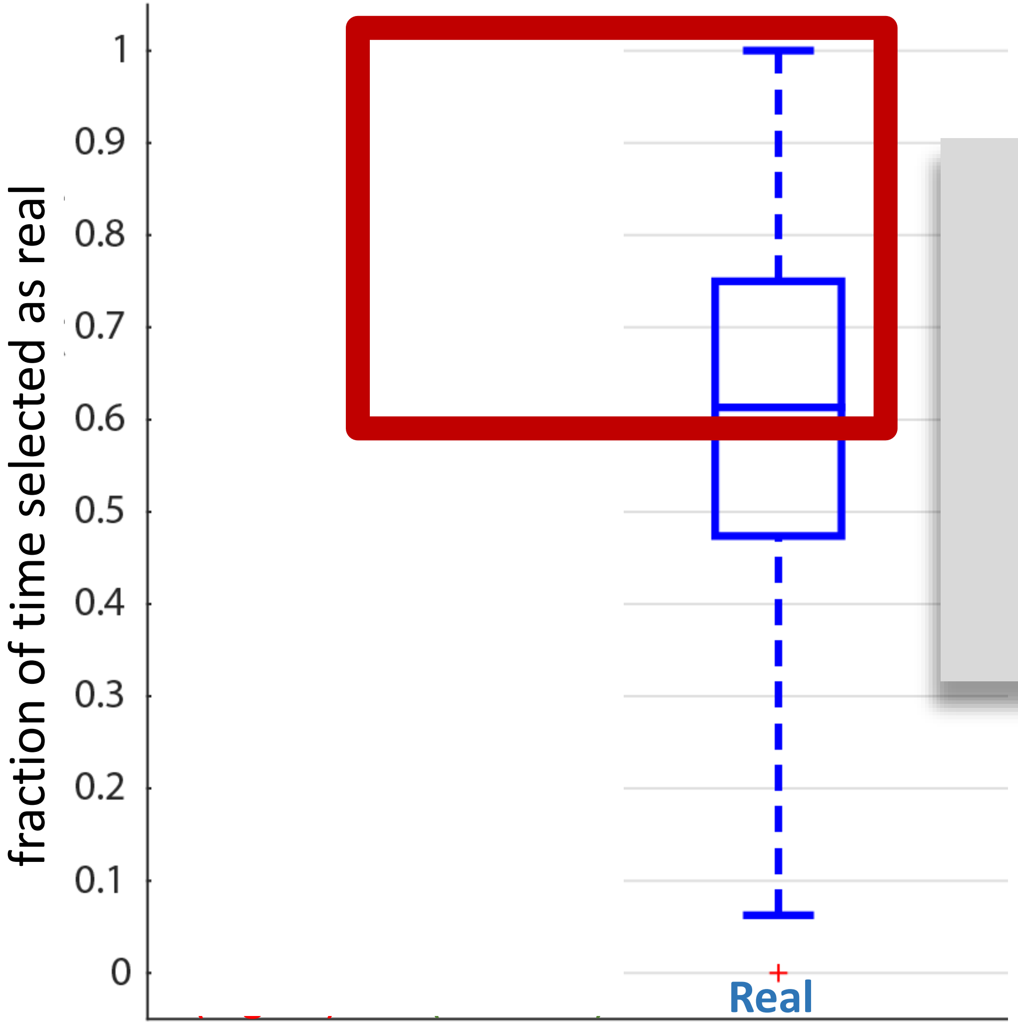

Ariel


ariel (0.9630)

Are Adversarial Eyeglasses Inconspicuous?



Are Adversarial Eyeglasses Inconspicuous?



Most realistic 10% of physically realized eyeglasses are more realistic than average real eyeglasses

Carnegie Mellon University CyLab Security and Privacy Institute

Can *an Attacker* Fool ML Classifiers? (Attempt #2)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects in a limited way

Can't control camera position, lighting

Defender / beholder doesn't notice attack (to be measured by user study)

Carnegie Mellon University CyLab Security and Privacy Institute

Considering Camera Position, Lighting

- Used algorithm to measure pose (pitch, roll, yaw)
- Mixed-effects logistic regression
  - Each 1° of yaw = 0.94x attack success rate
  - Each 1° of pitch = 0.94x (VGG) or 1.12x (OpenFace) attack success rate

- Varied luminance
  (add 150W incandescent light at 45°, 5 luminance levels)
  - Not included in training → 50% degradation in attack success
  - Included in training → no degradation in attack success

Carnegie Mellon University CyLab Security and Privacy Institute

What If Defenses Are in Place?

- Already:
  - Augmentation to make face recognition more robust to eyeglasses
- New:
  - Train attack detector (Metzen et al. 2017)
    - 100% recall and 100% precision
  - Attack must fool original DNN and detector
- **Result** (digital environment): **attack success unchanged**, with minor impact to conspicuousness

Can *an Attacker* Fool ML Classifiers? (Attempt #2)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects in a limited way ✓

Can't control camera position, lighting ✓ ?

Defender / beholder doesn't notice attack
(to be measured by user study) ✓

Other Attack Scenarios?

Dodging: One pair of eyeglasses, many attackers?

Change to training process:

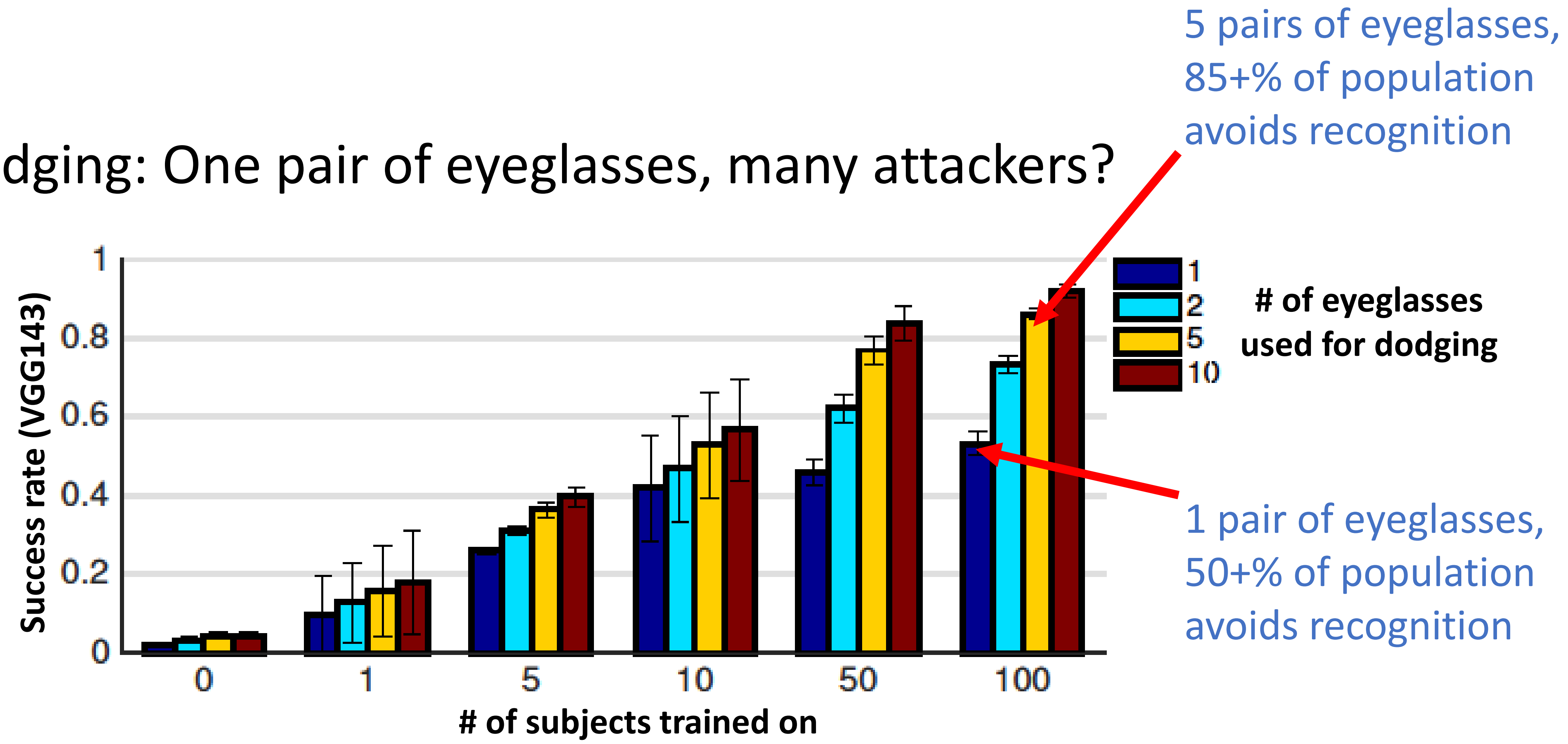
Train with multiple images of one user
$\longrightarrow$ train with multiple images of *many* users

Create multiple eyeglasses, test with large population

Carnegie Mellon University CyLab Security and Privacy Institute

Other Attack Scenarios?

Dodging: One pair of eyeglasses, many attackers?



Other Attack Scenarios?
or Defense

Privacy protection?

- E.g., against mass surveillance at a political protest

Unhappy speculation: individually, probably not

- 90% of video frames successfully misclassified
  → 100% success at defeating laptop face logon
  → 0% at avoiding being recognized at a political protest

Carnegie Mellon University CyLab Security and Privacy Institute

Other Attack Scenarios?
or Defense

Denial of service / resource exhaustion:

"appear" in many locations at once,
e.g., for surveillance targets to evade pursuit

Other Attack *or Defense* Scenarios?

Stop sign ⟶ speed limit sign [Eykholt et al., arXiv ‘18]

Other Attack Scenarios?
or Defense

Stop sign → speed limit sign [Eykholt et al., arXiv ’18]

Hidden voice commands [Carlini et al., ’16-19]

noise → “OK, Google, browse to evil dot com”

Malware classification [Suciu et al., arXiv ’18]

malware → “benign”

Fooling ML Classifiers: Summary and Takeaways

- “Attacks” may not be meaningful until we fix context
  - E.g., for face recognition:
    - Attacker: physically realized (i.e., constrained) attack
    - Defender / observer: attack isn’t noticed as such
- Even in a practical (constrained) context, real attacks exist
  - Relatively robust, inconspicuous; high success rates
- Hard-to-formalize constraints can be captured by a DNN
- Similar principles about constrained context apply to other domains: e.g., malware, spam detection

For more: www.ece.cmu.edu/~lbauer/proj/advml.php

Carnegie Mellon University
CyLab
Security and Privacy Institute