Real and Stealthy Attacks on State-of-the-Art Face Recognition *

Mahmood Sharif, Sruti Bhagavatula, and Lujo Bauer, CMU; Michael K. Reiter, UNC

Introduction

Machine learning (ML) is ubiquitous, enables revolutionary technologies:

If ML fails:

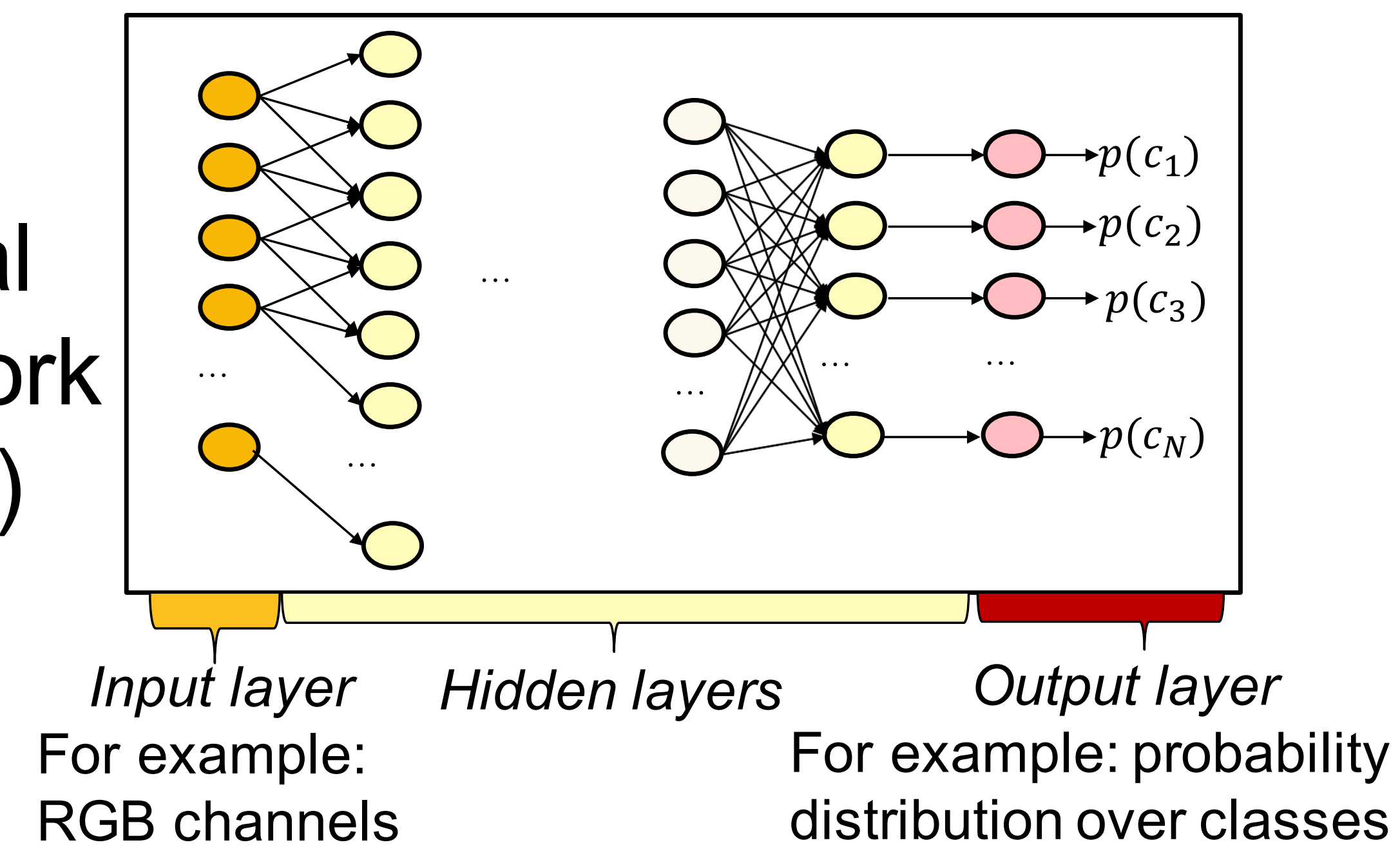
Our research questions investigate robustness of ML algorithms:

Can attackers make ML fail? Can attacks be *inconspicuous* and *physically realizable?*

Background and Prior Work

ML classifiers (e.g., in intrusion detection, cancer detection, …) are functions from inputs to classes (or probability distributions over classes)

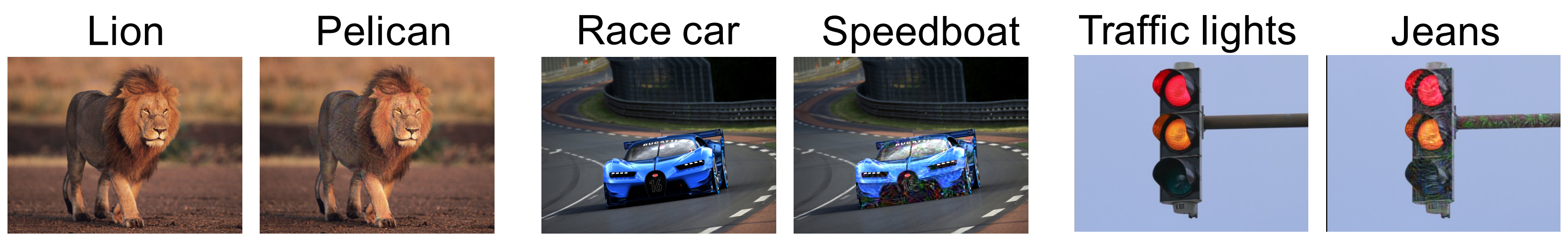
Deep Neural Network (DNN)


$p(c_1)$
$p(c_2)$
$p(c_3)$
$p(c_N)$
Input layer
For example: RGB channels
Hidden layers
Output layer
For example: probability distribution over classes

Imperceptible attacks have been demonstrated that confuse deep neural networks (DNNs) [1], by solving:

$$argmin_r \underbrace{|f(x+\mathbf{r})-l|}_{\text{misclassification}} + c\underbrace{|\mathbf{r}|}_{\text{imperceptibility}}$$

x is the input image; $f(\cdot)$ is the classification function (e.g., DNN); l is the desired output class; r: perturbation (or change applied to the input).



Our Approach and Results

Our focus: DNNs for state-of-the-art face recognition [2]

Attack goals:

- *Impersonation*: being classified as specific target
- *Dodging*: not being classified as self

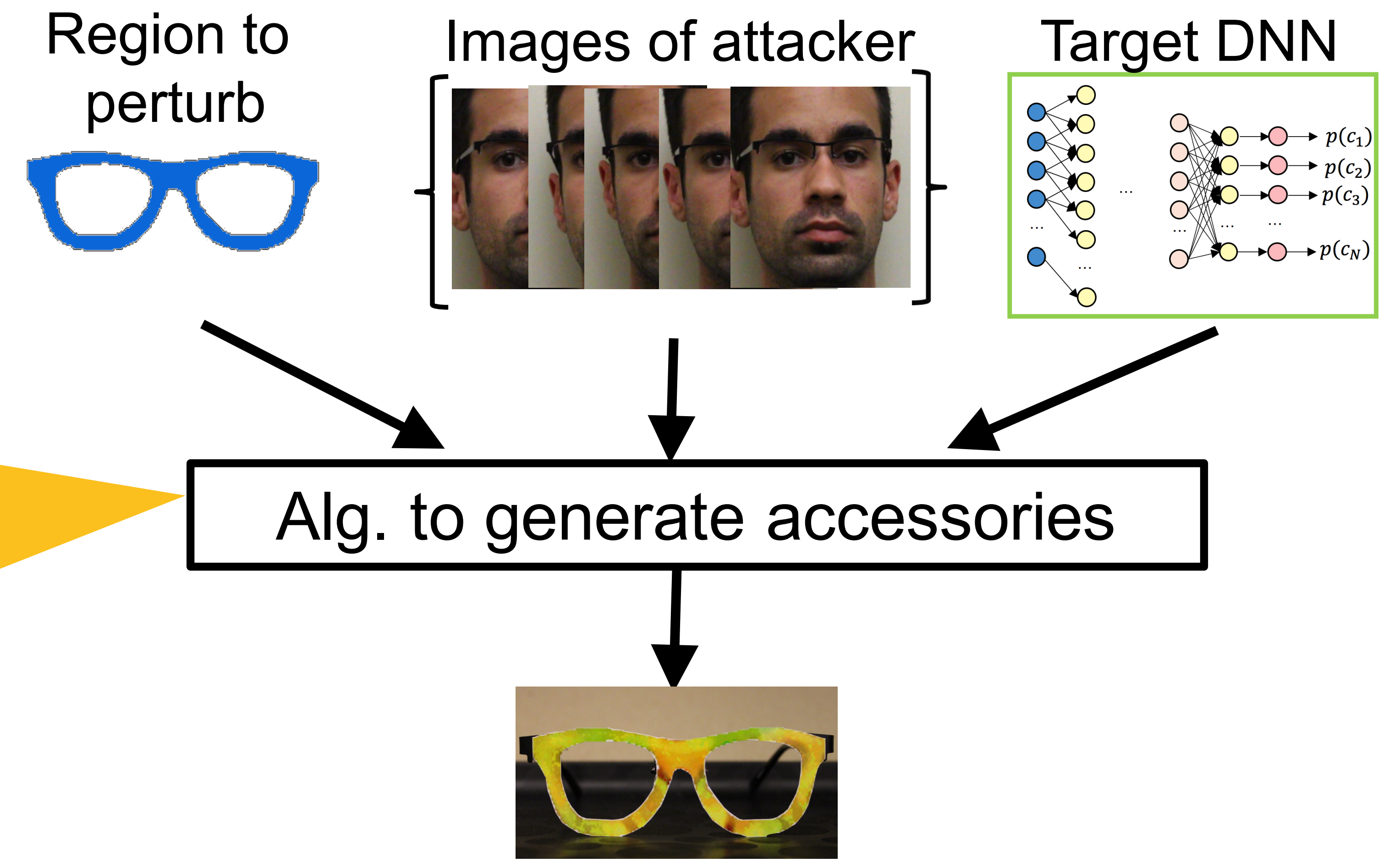
We create realizable, inconspicuous attacks by:

1. Limiting perturbation to eyeglass frames
2. Minimizing total variations (TV) btw. adjacent pixels
3. Minimizing "non-printability score" (NPS)
4. Increasing robustness: an attack should fool the system for more than one face image

Objective for impersonation (dodging is analogous):

$$\underset{r}{\text{argmin}}\left(\left(\sum_{\text{x}\in\text{X}}|f(x+\mathbf{r})-l|\right)+\kappa_1 TV(r)+\kappa_2 NPS(r)\right)$$

Attack generation:



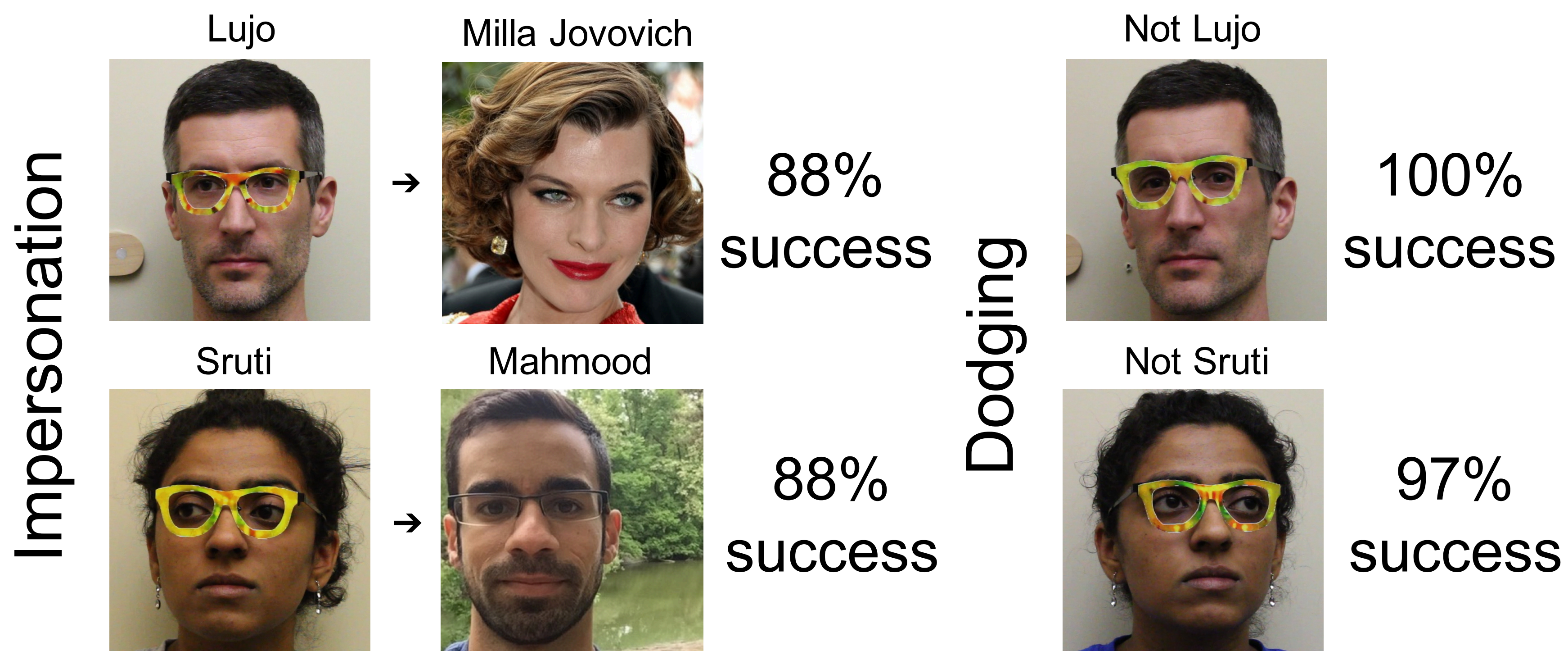
Region to perturb

Images of attacker

Target DNN

Alg. to generate accessories

Results: fool DNN trained on 7 subjects + 3 authors



Lujo → Milla Jovovich: 88% success

Sruti → Mahmood: 88% success

Not Lujo: 100% success

Not Sruti: 97% success

In paper: more experiments with larger DNN*

* M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *CCS*, 2016.

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *BMVC*, 2015.

Carnegie Mellon University CyLab Security and Privacy Institute