# Efficient Decision-based Black-box Adversarial Attacks on Face Recognition


Yinpeng Dong[1], Hang Su[1], Baoyuan Wu[2], Zhifeng Li[2], Wei Liu[2], Tong Zhang[3], Jun Zhu[1]

[1] Institute for AI, Tsinghua University, [2] Tencent AI Lab, [3] Hong Kong University of Science and Technology

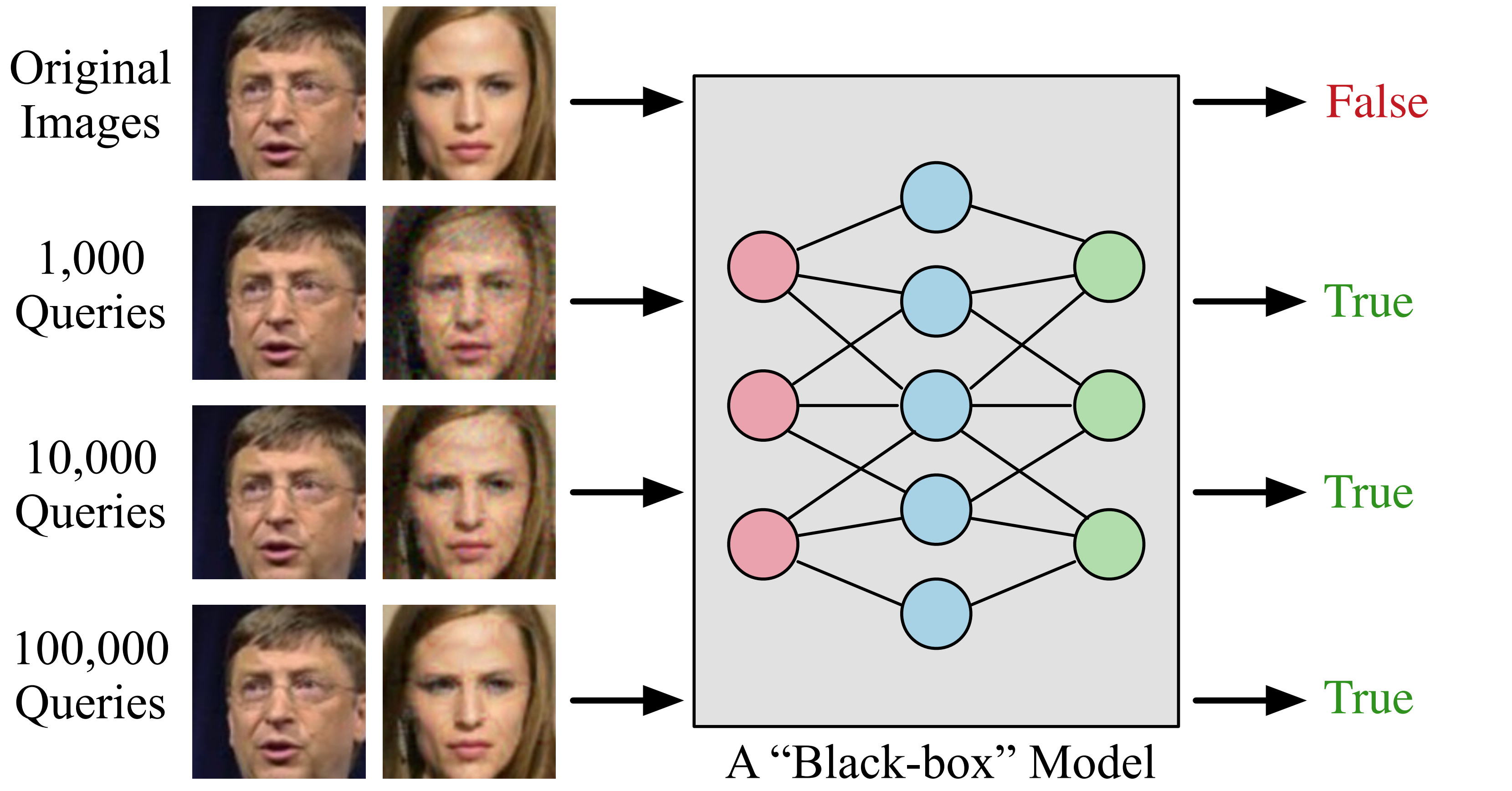
Tsinghua University   Tencent AI Lab



CVPR LONG BEACH CALIFORNIA June 16-20, 2019


## Introduction

- **Face recognition** models based on deep neural networks are vulnerable to **adversarial examples**. In many real-world face recognition applications, the attackers cannot get access to the model details.
- We focus on the realistic **decision-based black-box setting**, where **no** model information is exposed except that the attackers can only **query** the target model and obtain corresponding **hard-label predictions**. We are the **first** to study adversarial attacks on face recognition in the black-box setting.
- **Goal: Finding minimum adversarial perturbations by limited queries.**




## Problem Formulation

- **Constrained optimization problem**

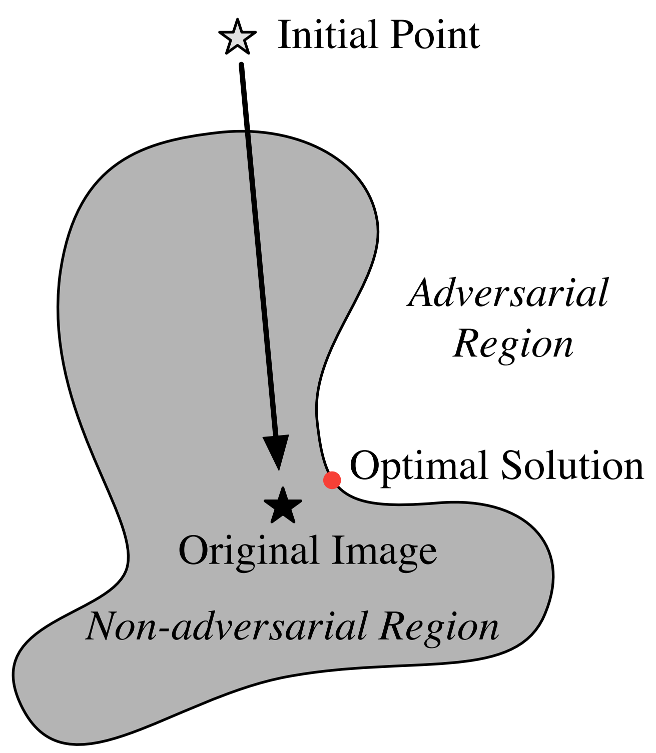
$$\min_{x^*} D(x^*, x), \quad s.t.\ C(f(x^*)) = 1$$

- $D$ is a distance metric (e.g., $L_2$ norm), $C$ is an adversarial criterion ($C(f(x)) = 0$).

- **A reformulation**

$$\min_{x^*} L(x^*) = D(x^*, x) + \delta(C(f(x^*)) = 1)$$

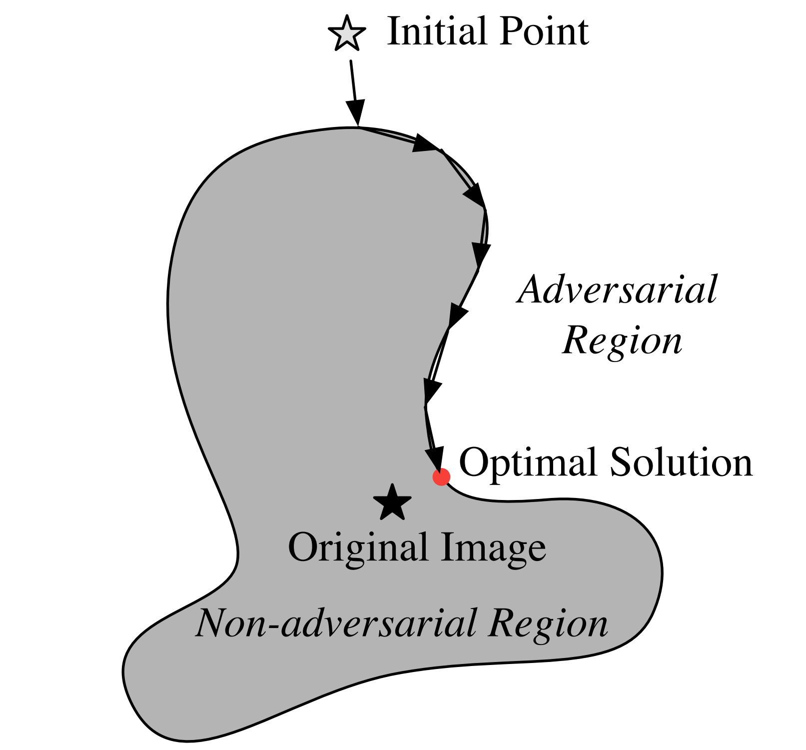
- $\delta(a) = 0$ if $a$ is true; otherwise $\delta(a) = +\infty$.




- **Dodging attack: protect personal privacy**
  - $C(f(x^*)) = \mathbb{I}(f(x^*) = 0)$ in face verification;
  - $C(f(x^*)) = \mathbb{I}(f(x^*) \neq y)$ in face identification.
- **Impersonation attack: evade face authentication systems**
  - $C(f(x^*)) = \mathbb{I}(f(x^*) = 1)$ in face verification;
  - $C(f(x^*)) = \mathbb{I}(f(x^*) = y^*)$ in face identification.

## Previous Methods

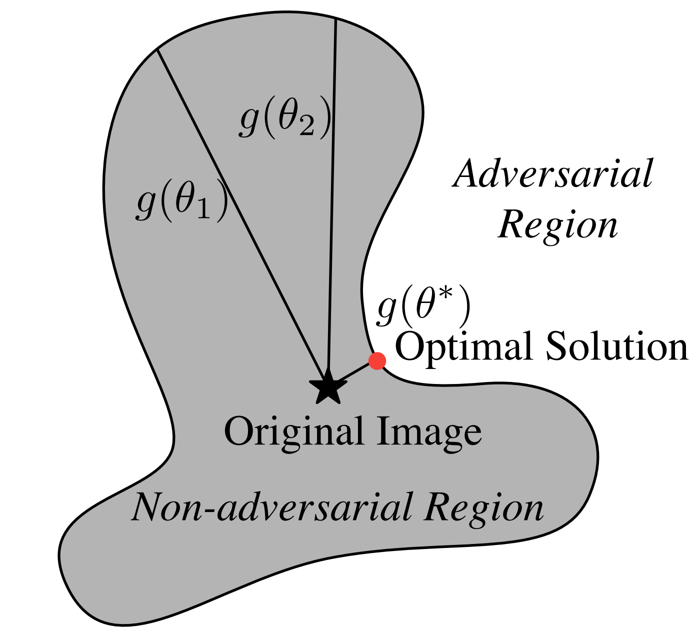
**Boundary Attack [Brendel et al., 2018]**




Random search on the decision boundary

**Optimization Attack [Cheng et al., 2019]**

$$g(\theta) = \operatorname*{argmin}_{\lambda > 0} \left( C\left( f\left( x + \lambda \frac{\theta}{\|\theta\|} \right) \right) = 1 \right)$$




Zeroth-order optimization to find optimal $\theta$

**But they usually require a tremendous number of queries (~$10^5$) to converge, or get a relatively large perturbation given a limited budget of queries.**

## Evolutionary Attack

- **The evolutionary attack can improve the query efficiency by modeling the local geometry of the search directions and reducing the dimension of the search space.**
- **Algorithm:**

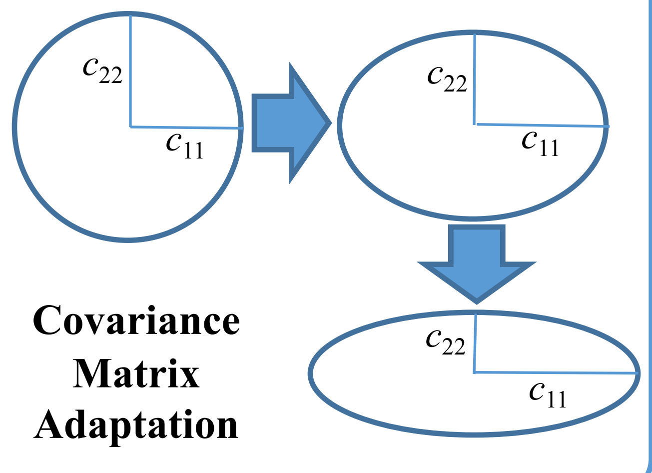
**Input:** original image $x$; the dimension $n$ of the input space, $m$ of the search space;

1. Initialize $C = I_m$, $p_c = 0$, $\tilde{x}^*$ as an adversarial example;
2. For $t = 1$ to $T$ do
3. Sample $z = N(0, \sigma^2 C)$;
4. Select $k$ coordinates with probability proportional to the diagonal element in $C$;
5. Set the non-selected elements to 0;
6. Upscale $z$ to $\mathbb{R}^n$ by bilinear interpolation and get $\tilde{z}$;
7. $\tilde{z} \leftarrow \tilde{z} + \mu(x - \tilde{x}^*)$;
8. If $L(\tilde{x}^* + \tilde{z}) < L(\tilde{x}^*)$ then
9. $\tilde{x}^* \leftarrow \tilde{x}^* + \tilde{z}$;
10. $p_c = (1 - c_c)p_c + \sqrt{c_c(2 - c_c)}\frac{z}{\sigma}$;
11. $c_{ii} = (1 - c_{cov})c_{ii} + c_{cov}(p_c)_i^2$;
12. End if
13. End for
14. Return $\tilde{x}^*$.

*Reduce the dimension of search space by dimensionality reduction and stochastic coordinate selection*

*Add a bias to reduce the distance $D(x^*, x)$*

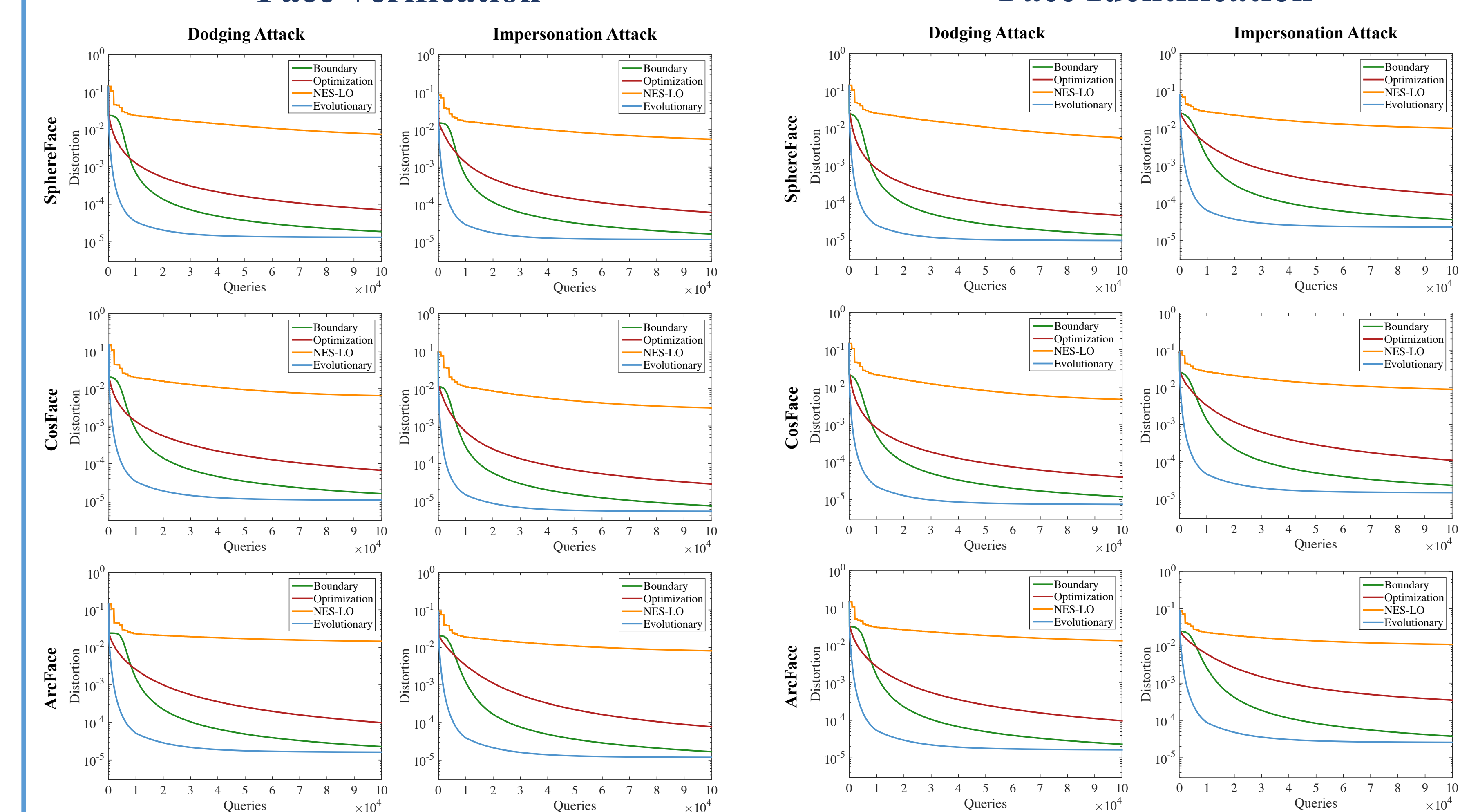
*Use a diagonal covariance matrix to model the local geometry of the search directions*

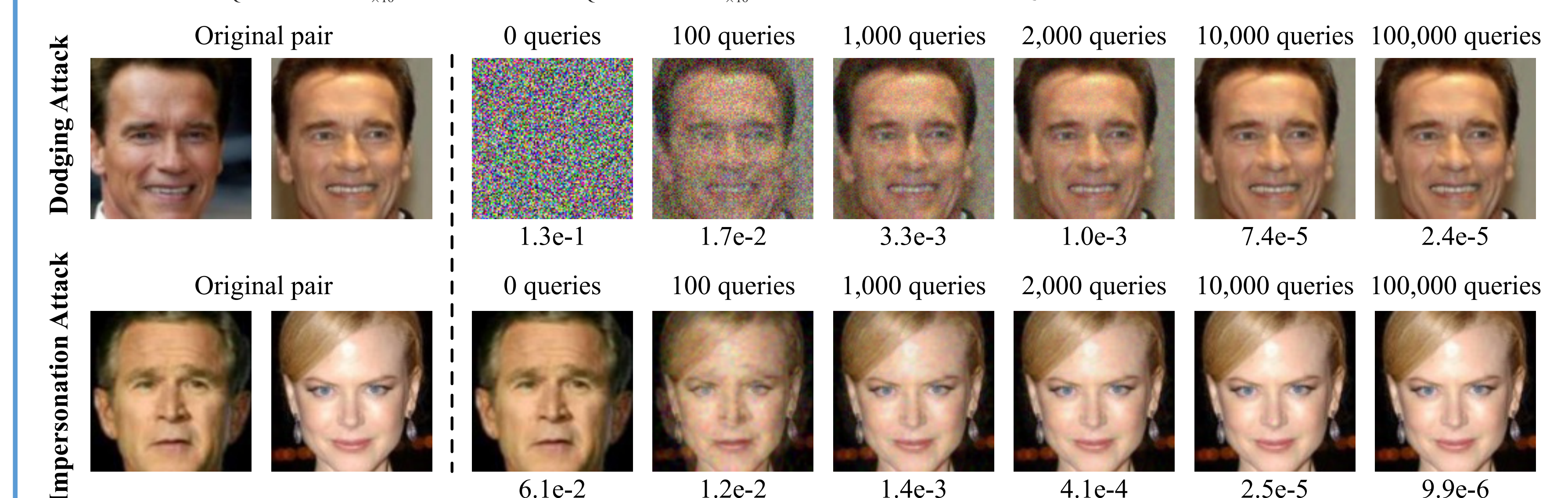



**Covariance Matrix Adaptation**

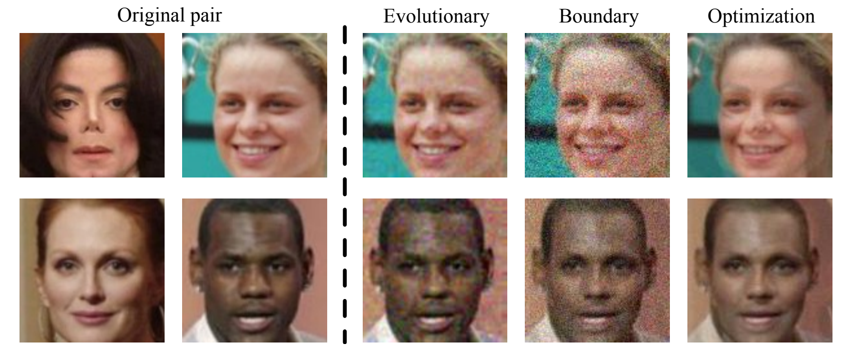
## Experiments

- **Attacks on SphereFace, CosFace, and ArcFace**




- **Attacks on the face verification API in Tencent AI Open Platform**




| Attack Method | Distortion (MSE) |
|---|---|
| Boundary [1] | 1.63e-2 |
| Optimization [4] | 1.71e-2 |
| Evolutionary | **2.54e-3** |

## Conclusion

- **We propose an evolutionary attack method to improve query efficiency in the decision-based black-box setting;**
- **We demonstrate the practical applicability by attacking a real-world face recognition system;**
- **Our attack can be used to protect personal privacy and evaluate the robustness of face recognition models.**


www.PosterPresentations.com