

Adversarial attacks on image classifiers

Andrea Cavallaro

IEEE SPS Distinguished Lecture - 29 October 2020

$$y = M(x)$$



training process



training data

Adversarial example

- Intentionally perturbed signal that **misleads (evades)** one or more **classifiers**

original (clean) signal

$\mathbf{x}$

$$\mathbf{y} = \mathbf{M}(\mathbf{x})$$

class
label

classifier

?

perturbed (adversarial) signal

$\dot{\mathbf{x}}$

$$\mathbf{y} \neq \mathbf{M}(\dot{\mathbf{x}})$$

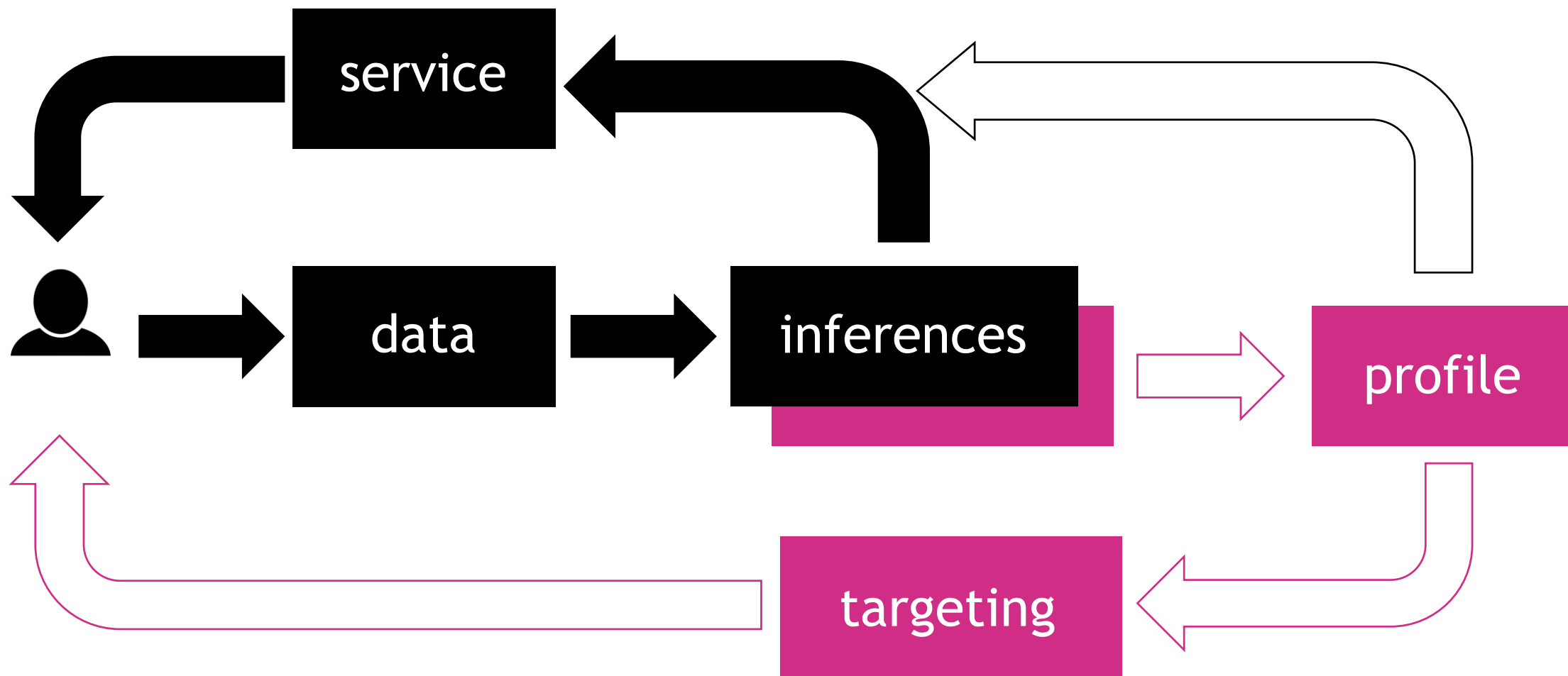
What is the problem?

optimization
for/on the user

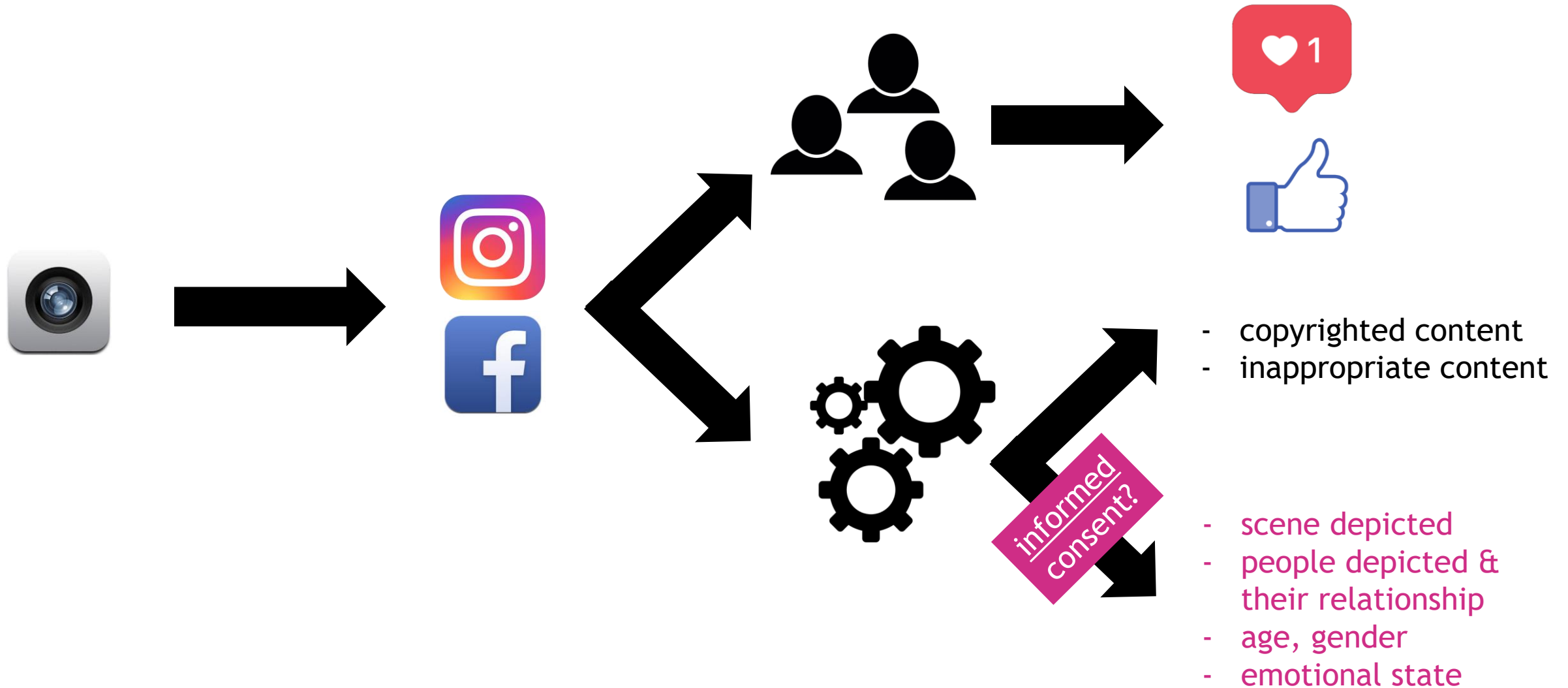
user as source and target
of a process

undesirable
inferences

defending from a 'machine'



Example: routine inferences on images shared in social media



Personal data and consent

Personal data: *“any information relating to an identified or identifiable natural person (**‘data subject’**)”*

Consent should be a *“freely given, specific, informed and unambiguous indication of the data subject’s wishes”*.

Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation)

Contrasting objectives?

informed
consent

data
minimization

*“Personal data shall be
adequate, relevant and limited
to what is necessary in relation
to the purposes for which they
are processed”*

vs

data
monetization

Learning data manipulations for privacy protection



to **protect** the private content of images
(that a user shares with other users)
from **undesirable** automatic inferences

Properties of a ‘good’ adversarial attack

effectiveness

success rate in
evading a classifier

targeted

untargeted

robustness

success rate in
evading a classifier
despite defenses

transferability

success rate in
evading an unseen classifier

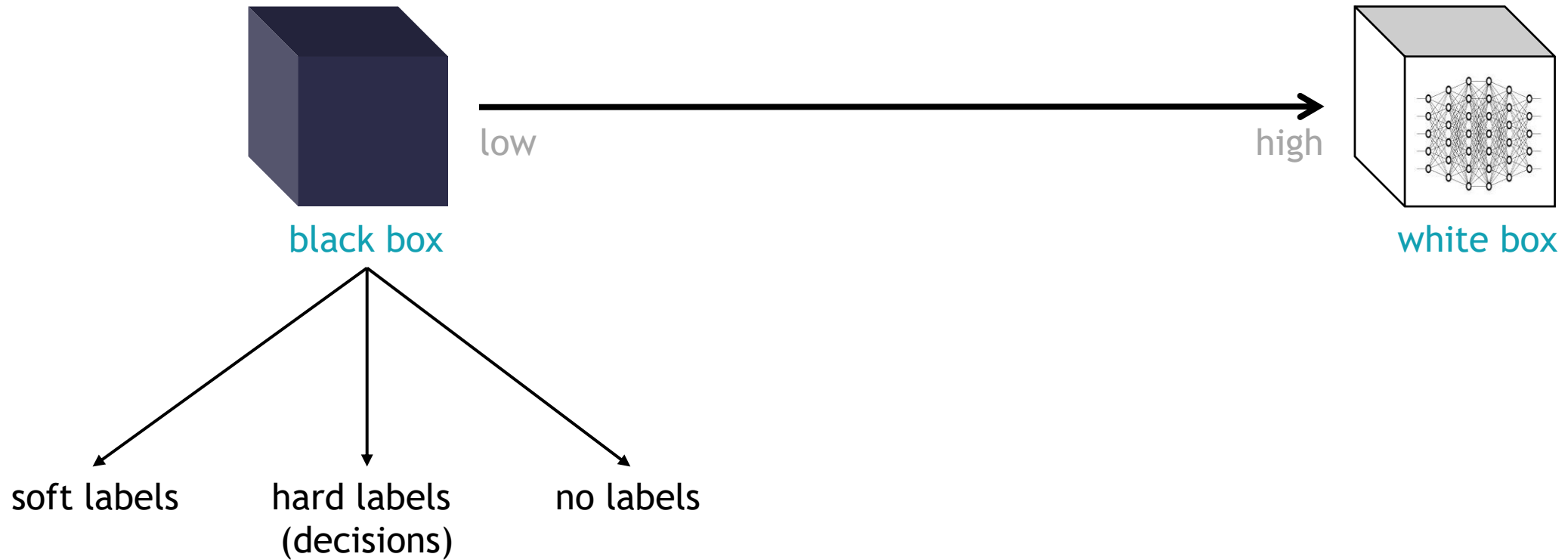
un-detectability

signal “perceived”
as an original one

irreversibility

if the manipulation is detected,
low probability of
recovering the original class

Knowledge about the classifier (or its output)



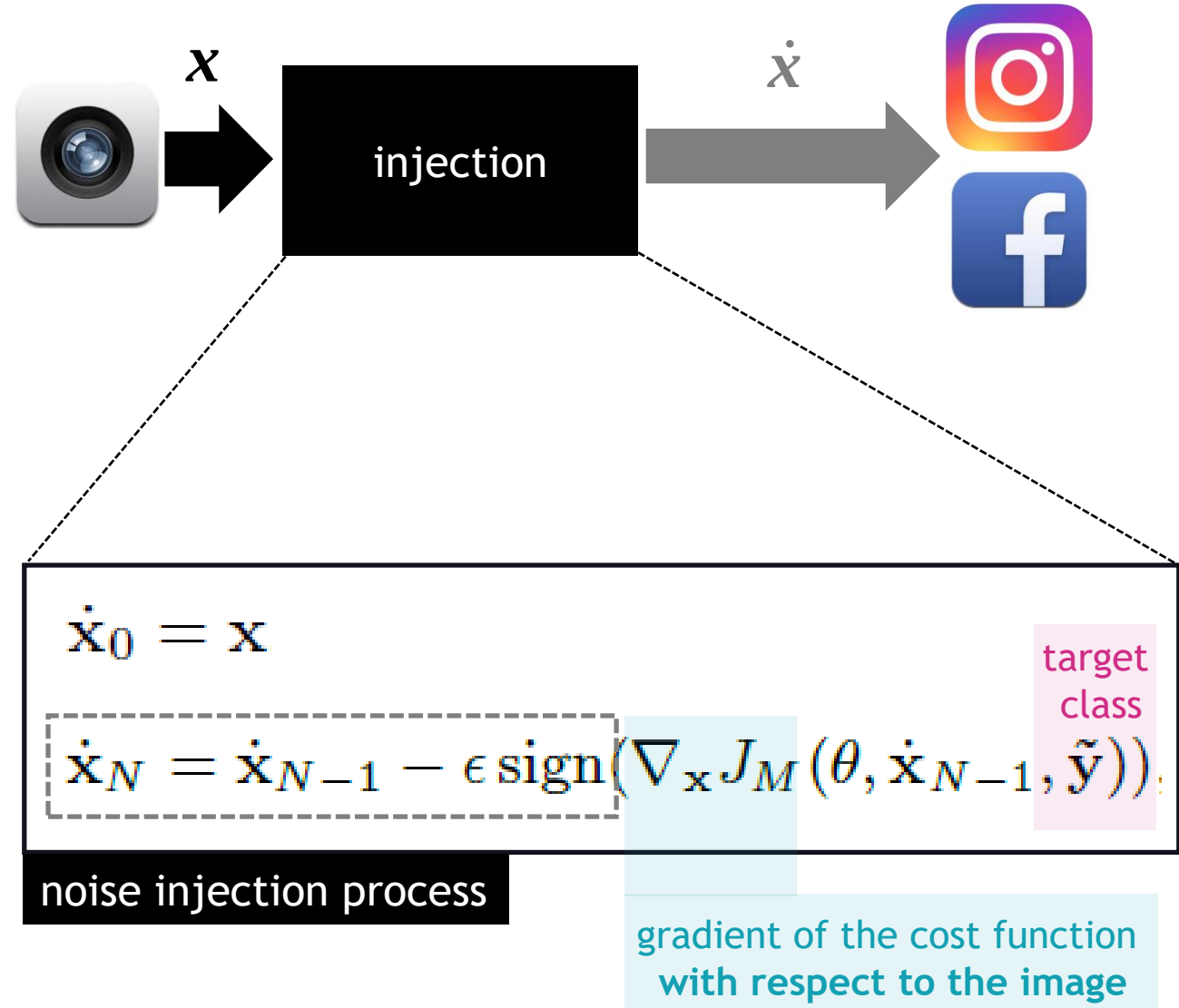
References

- R.S. Matilla, C.Y. Li, A.S. Shamsabadi, R. Mazzon, A. Cavallaro. **Exploiting vulnerabilities of deep neural networks for privacy protection.** *IEEE Transactions on Multimedia*, 22 (7), doi: 10.1109/TMM.2020.2987694, 2020.
- A. Modas, R.S. Sanchez-Matilla, P. Frossard, A. Cavallaro. **Towards robust sensing for autonomous vehicles.** *IEEE Signal Processing Magazine*, 37 (4), doi: 10.1109/MSP.2020.2985363, 2020.
- M.S. Cross, A. Cavallaro. **Privacy as a feature for body worn cameras.** *IEEE Signal Processing Magazine*, 37 (4), doi: 10.1109/MSP.2020.2989686, 2020.
- G. Tadesse, O. Bent, L. Marcenaro, K. Weldemariam, A. Cavallaro. **Privacy-aware human activity recognition from a wearable camera.** *IEEE Signal Processing Magazine*, 37 (3), doi: 10.1109/MSP.2020.2976158, 2020.
- A.S. Shamsabadi, R.S. Matilla, A. Cavallaro. **ColorFool: Semantic adversarial colorization.** *In Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, USA, 2020.
- A.S. Shamsabadi, C. Oh, A. Cavallaro. **EdgeFool: An adversarial image enhancement filter.** *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.
- C. Y. Li, A. S. Shamsabadi, R. S. Matilla, R. Mazzon, A. Cavallaro. **Scene privacy protection.** *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.
- O. Sarwar, B. Rinner, A. Cavallaro. **A privacy-preserving filter for oblique face images based on adaptive hopping Gaussian mixtures.** *IEEE Access*, 7, doi: 10.1109/ACCESS.2019.2944861, 2019.
- O. Sarwar, A. Cavallaro, B. Rinner. **Temporally smooth privacy protected airborne videos.** *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018.

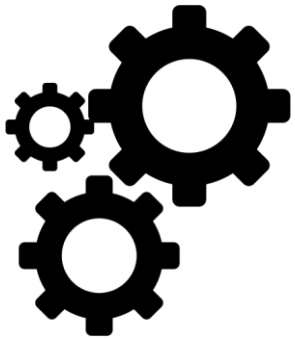
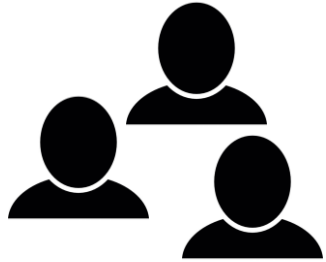
P-FGSM (Private Fast Gradient Sign Method)

- Approach

- perturbations constrained by ϵ
 - maintains visual quality
- **targeted** approach
 - adaptive subset that most likely does not include the class to be protected
 - favors irreversibility
- iterate, until
 - **desired classification probability** *or*
 - max number of iterations



Example



x



$$y = M(x)$$

class: church

83%

$\dot{x}$



$$y \neq M(\dot{x})$$

class: zen garden

99%

Properties of a ‘good’ adversarial attack: P-FGSM

effectiveness

success rate in
evading a classifier

robustness

success rate in
evading a classifier
despite defenses

transferability

success rate in
evading an unseen classifier

un-detectability

signal “perceived”
as an original one

irreversibility

if the manipulation is detected,
low probability of
recovering the original class

RP-FGSM (Robust and Private Fast Gradient Sign Method)

- Targeted or untargeted iterative approach
 - *objective*: to mislead seen & unseen classifiers with & without known defenses

$$\dot{\mathbf{x}}_{n+1} = \mathcal{C}_{\mathbf{x}, \epsilon} (\dot{\mathbf{x}}_n + \delta \dot{\mathbf{x}}_n)$$

$$\delta \dot{\mathbf{x}}_n = \begin{cases} \delta \operatorname{sign}(\nabla_{\mathbf{x}} J_{M_r}(\theta, \phi_r(\dot{\mathbf{x}}_n), y_{\mathbf{x}}^r)) \\ -\delta \operatorname{sign}(\nabla_{\mathbf{x}} J_{M_r}(\theta, \phi_r(\dot{\mathbf{x}}_n), y_t)) \end{cases}$$

randomly
selected
classifier

randomly
selected
defense

number of classifiers

class prediction probability

$$y_t = R \left(\bigcap_{k=1}^K \left\{ y_{j+1} : \sum_{i=1}^j p'_i{}^k > \gamma \right\} \right)$$

random selection

subset of classes whose cumulative probability exceeds a threshold

$$j \in \{1, \dots, D-1\}$$

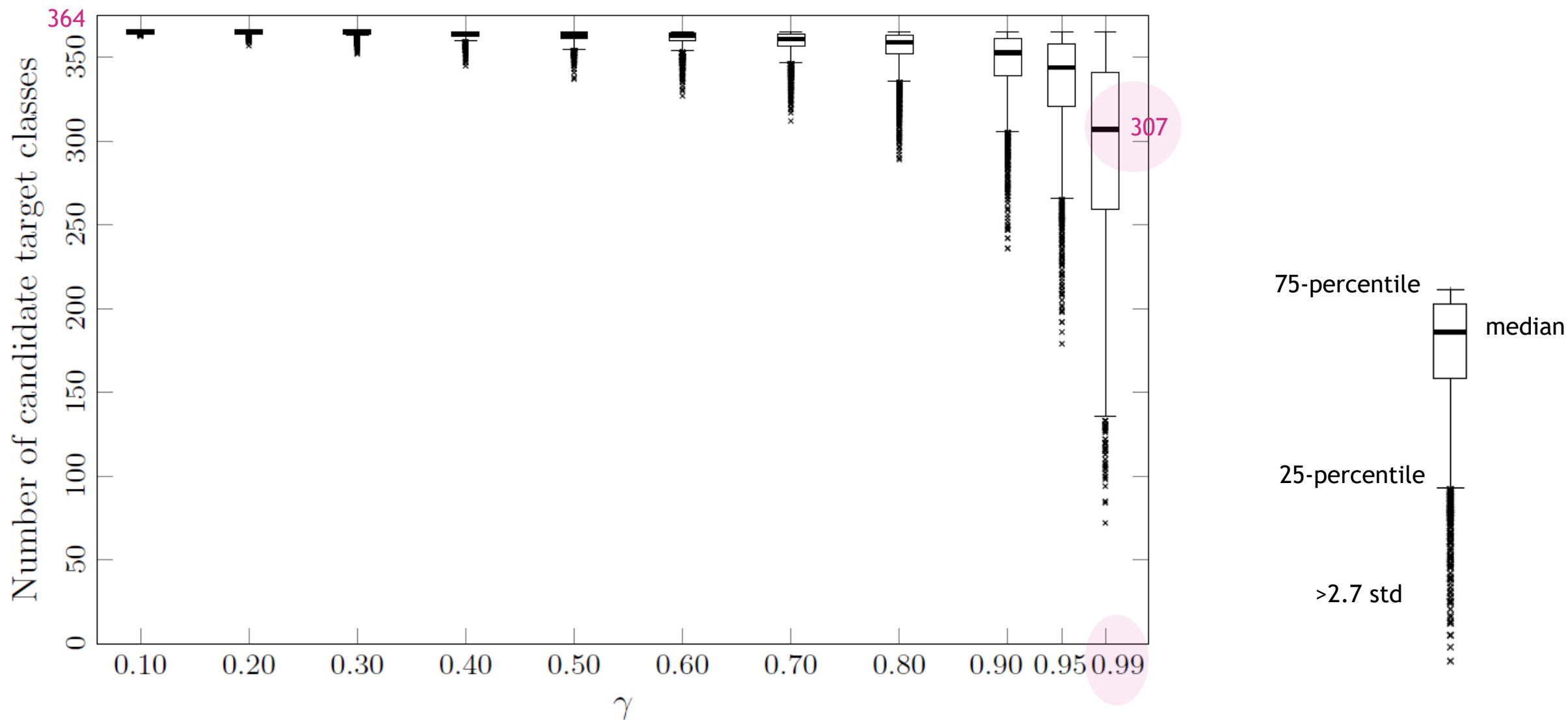
number of classes

Exploiting vulnerabilities of deep neural networks for privacy protection

R. Sanchez-Matilla, C.Y. Li, A.S. Shamsabadi, R. Mazzon, A. Cavallaro

IEEE Transactions on Multimedia, 22 (7), 2020

Number of classes to (randomly) choose the target from



Original images & their estimated class



railroad
track



water
park



airplane
cabin



church



bus
station



picnic
area



swimming
pool



waiting
room



sandbox

RP-FGMS adversarial images & their estimated class



Amount of perturbation to create adversarial examples

$$\dot{x} = x + \delta$$

constrained

- small Lp-norm perturbations (bounded)
- limited success with unseen classifiers
- vulnerable to defenses
(high-frequency spatial perturbations are easily defeated by **denoising** algorithms)

unconstrained

- more transferable to unseen classifiers
- more robust to defenses (less detectable)
- may severely **degrade** images
(large perturbations are noticeable)

x

Constrained perturbations: examples



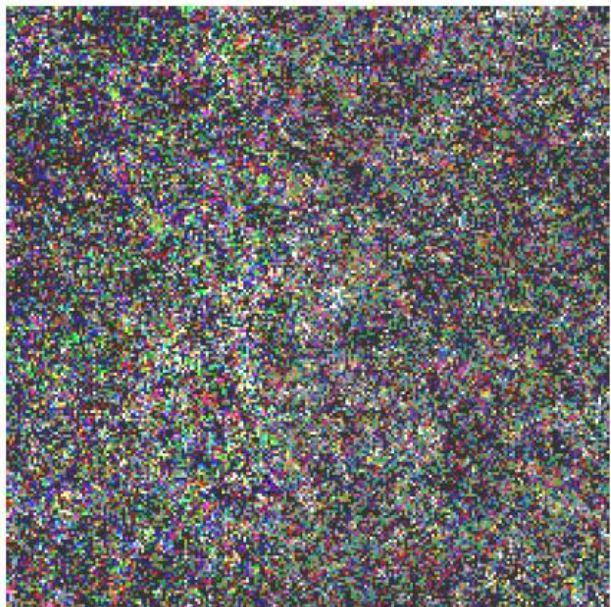
Original

δ

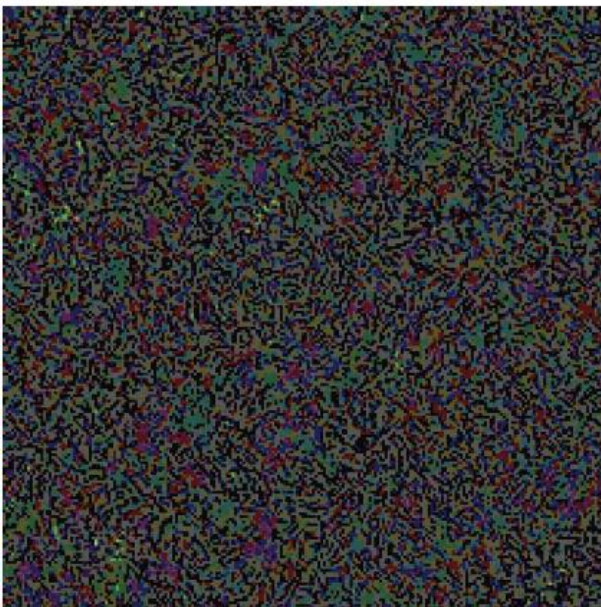
δ

δ

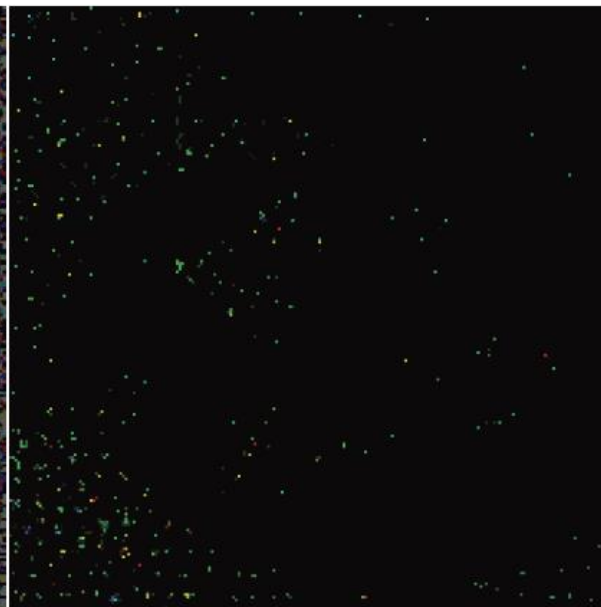
δ



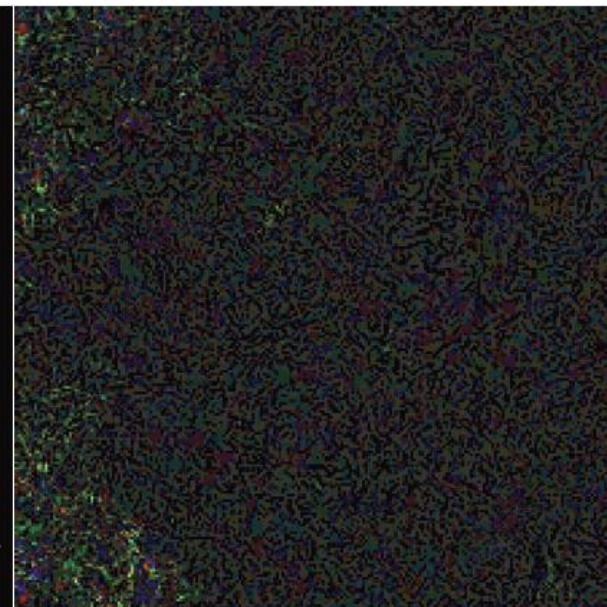
P-FGSM



DeepFool



SparseFool



CW

Unconstrained perturbations: approaches

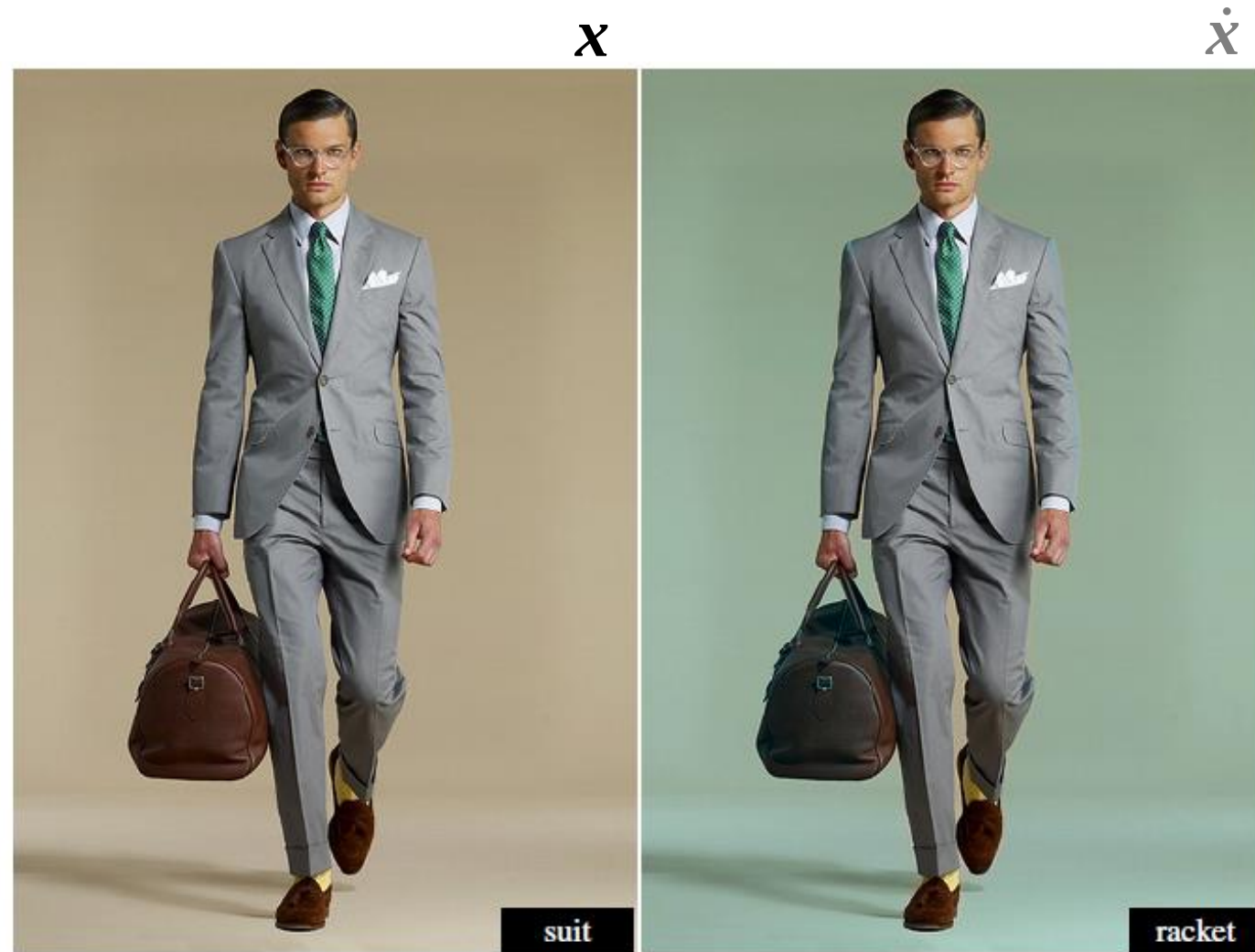
- **Shifting** randomly hue and saturation values
- Transferring **new textures** to images
- **Colorization**



can we do
any better?

ColorFool: content-based black-box adversarial attack

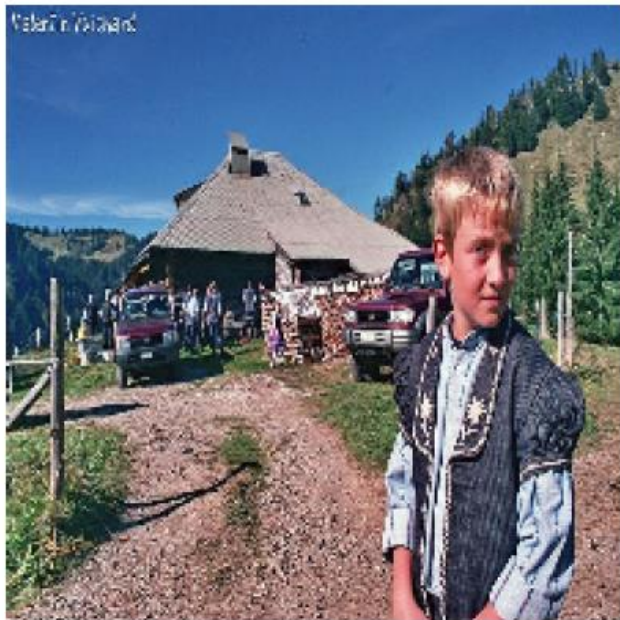
- Approach
 - **unconstrained** perturbations
 - selectively modifies colors within chosen ranges
 - maintains **natural colors**
 - uses Lab color space
 - exploits image semantics
 - **object categories**
 - person, vegetation, sky, water, other



ColorFool: semantic adversarial colorization

Shamsabadi, Sanchez-Matilla, Cavallaro

CVPR 2020





(any size)

Semantic
segmentation

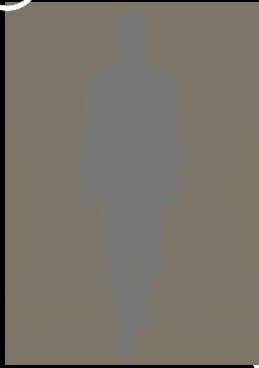


Color perturbation

RGB2Lab



Color perturbation



Lab2RGB



Let's test it!



Adversarial or original?

ColorFool-ed images



Original images



ColorFool-ed images



Original images



Do colors in the image look natural? [No | Yes]



150 Images

Opinion Score (Mean $\pm$ Std) for 39 people

Original

0.88 $\pm$ 0.33

SemanticAdv

0.22 $\pm$ 0.42

ColorFool

0.74 $\pm$ 0.44

Can I improve an image and turn it adversarial ?

EdgeFool: adversarial image enhancement filter

- Approach
 - learns structure-aware perturbations
 - end-to-end training
 - multi-task loss function accounts for
 - image detail enhancement *objective*
 - misleading *objective*



original (clean) image



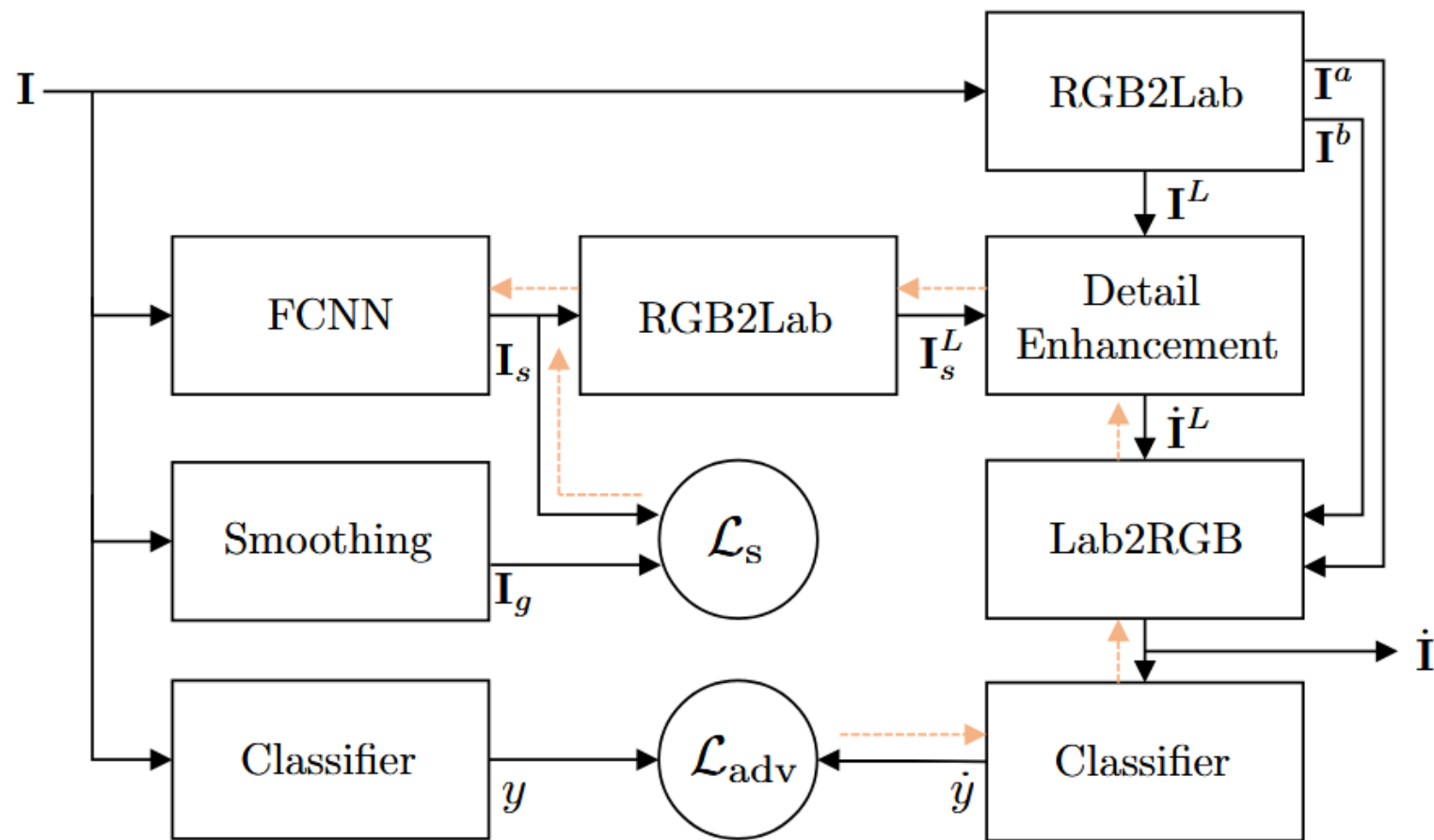
perturbed (adversarial) image

EdgeFool: an adversarial image enhancement filter

Shamsabadi, Oh, Cavallaro

IEEE ICASSP 2020

End-to-end training process



smoothing

$$\|\mathbf{I}_s - \mathbf{I}_g\|^2$$

$$\mathcal{L} = \alpha \mathcal{L}_s(\mathbf{I}_s, \mathbf{I}_g) + \mathcal{L}_{adv}(\dot{\mathbf{I}}, \mathbf{I})$$

adversarial loss

$$\dot{z}_y - \max\{\dot{z}_i : i = 1, \dots, D; i \neq y\}$$

logit score of class i

Original images



EdgeFool-ed images



Code

Exploiting vulnerabilities of deep neural networks for privacy protection

IEEE Transactions on Multimedia, 22 (7), 2020

R.S. Matilla, C.Y. Li, A.S. Shamsabadi, R. Mazzon, A. Cavallaro

<https://github.com/smartcameras/RP-FGSM>

ColorFool: Semantic adversarial colorization

IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Seattle, Washington, USA, 2020

A.S. Shamsabadi, R.S. Matilla, A. Cavallaro

<https://github.com/smartcameras/ColorFool>

EdgeFool: An adversarial image enhancement filter

IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020

A.S. Shamsabadi, C. Oh, A. Cavallaro

<https://github.com/smartcameras/EdgeFool>

Scene privacy protection

IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

C.Y. Li, A.S. Shamsabadi, R.S. Matilla, R. Mazzon, A. Cavallaro

<https://github.com/smartcameras/P-FGSM>

Other privacy-preserving perturbations

- Privacy-preserving **drone** videography
 - *objective*: to robustly protect the identity, while limiting spatio-temporal distortions
 - **adaptively** distorts the face appearance as a function of its resolution
 - **locally** changes parameters values to prevent estimation attacks

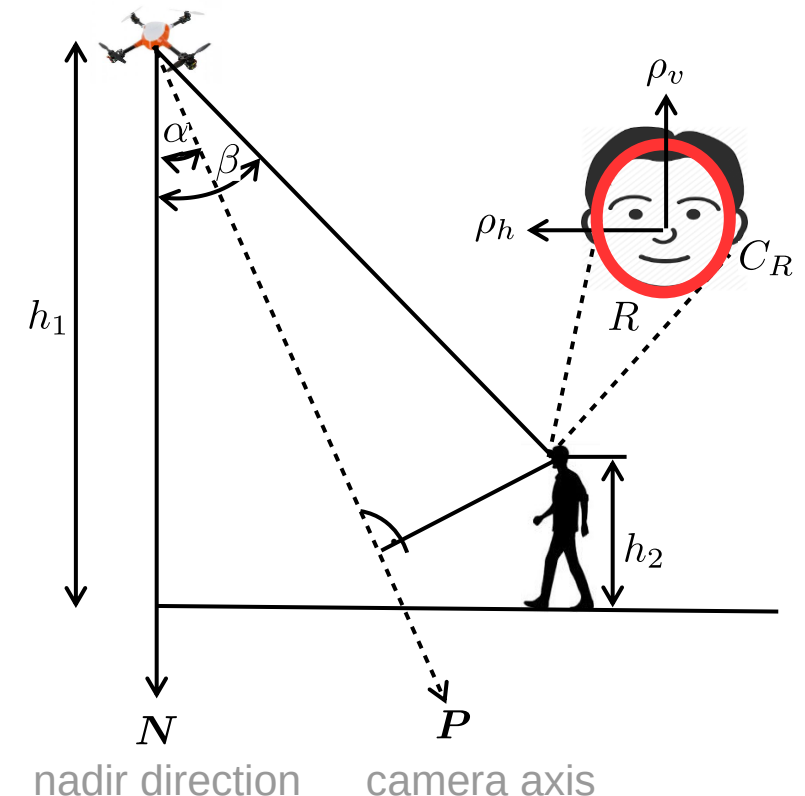


Temporally smooth privacy protected airborne videos

Sarwar, Cavallaro, Rinner
IROS 2018

A privacy-preserving filter for oblique face images
based on adaptive hopping Gaussian mixtures

Sarwar, Rinner, Cavallaro
IEEE Access, October 2019



Physical adversarial attacks

- Perturbations that modify the **environment** in an adversarial manner

to generate
observations
of inexistent objects

to hide
objects

to induce
misclassification
of detected objects

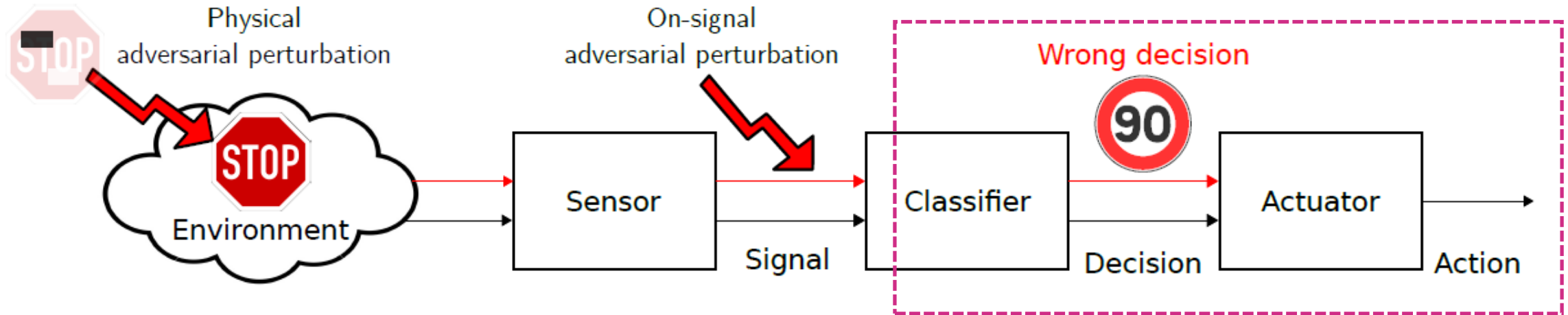
Towards robust sensing for autonomous vehicles

Modas, Sanchez-Matilla, Frossard, Cavallaro

IEEE Signal Processing Magazine, July 2020

Adversarial attacks on autonomous systems

- Safety critical application



Towards robust sensing for autonomous vehicles

Modas, Sanchez-Matilla, Frossard, Cavallaro

IEEE Signal Processing Magazine, July 2020

Outlook

$$\mathbf{y} = \mathbf{M}(\mathbf{x})$$

↑
training process

↑
training data

evaluation

fusion

adversarial
training

robustness

Towards robust sensing for autonomous vehicles

Modas, Sanchez-Matilla, Frossard, Cavallaro

IEEE Signal Processing Magazine, July 2020

Adversarial attacks on image classifiers

Andrea Cavallaro
a.cavallaro@qmul.ac.uk

Many thanks to
Janice Li, Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla,
Riccardo Mazzon, Changjae Oh, Omair Sarwar, Apostolos Modas

This work was supported by the Alan Turing Institute (EP/N510129/1), which is funded by the EPSRC, through the project PRIMULA; and by the CHIST-ERA project CORSMAL, under UK EPSRC grant EP/S031715/1.