ATTACKING DEEP NEURAL NETWORKS WITH ADVERSARIAL IMAGES

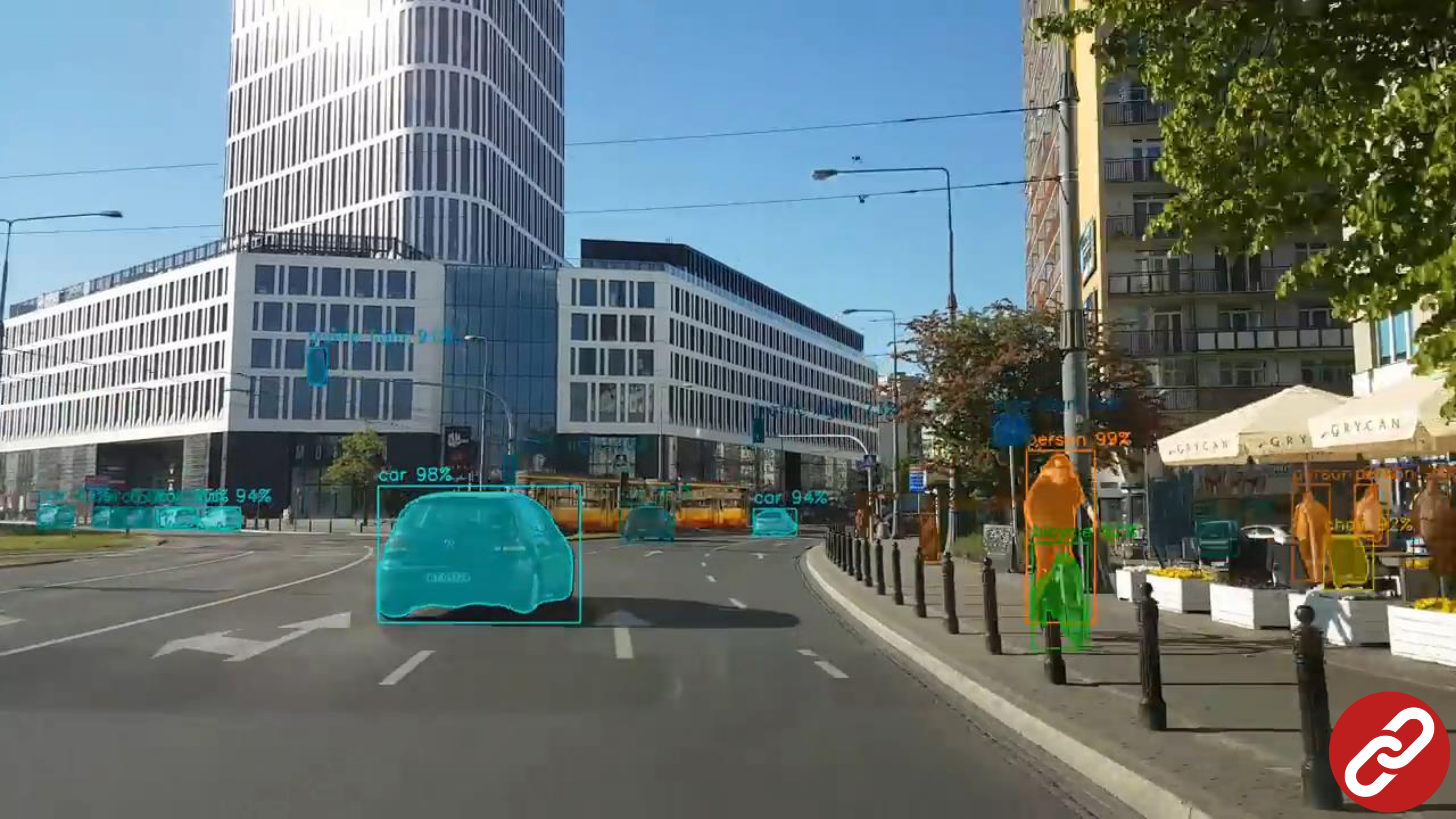
Fabrizio Falchi

ISTI, CNR, Pisa, italy

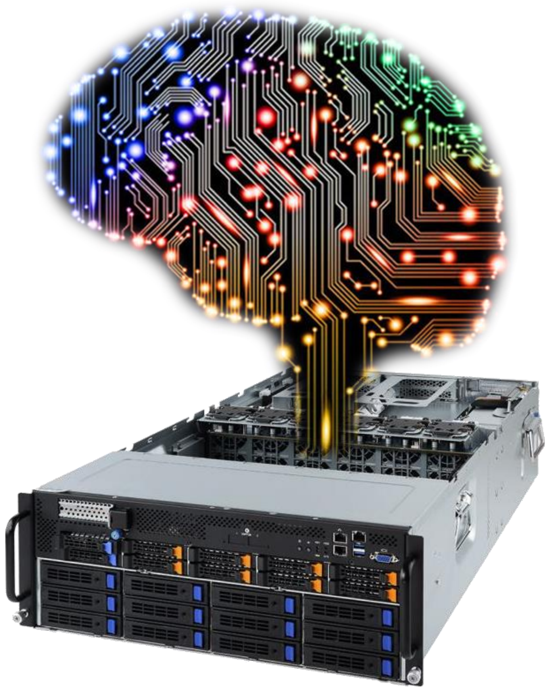
www.fabriziofalchi.it

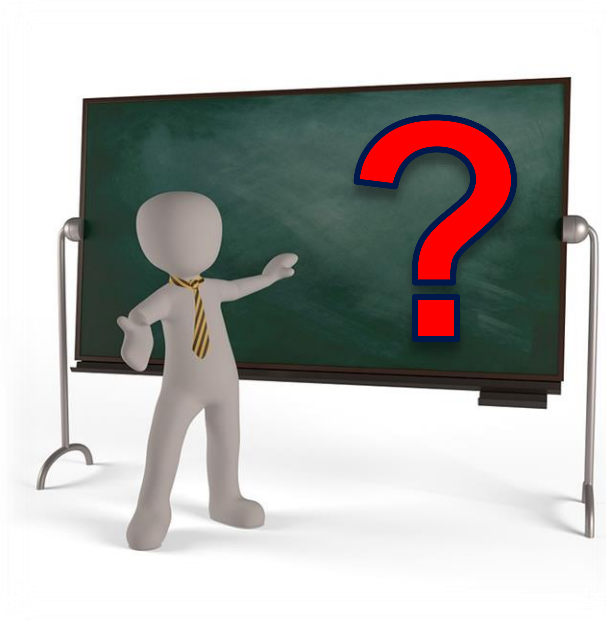
COST ACTION CA16101 - Dubrovnik, November 7th


TOYOTA HSR

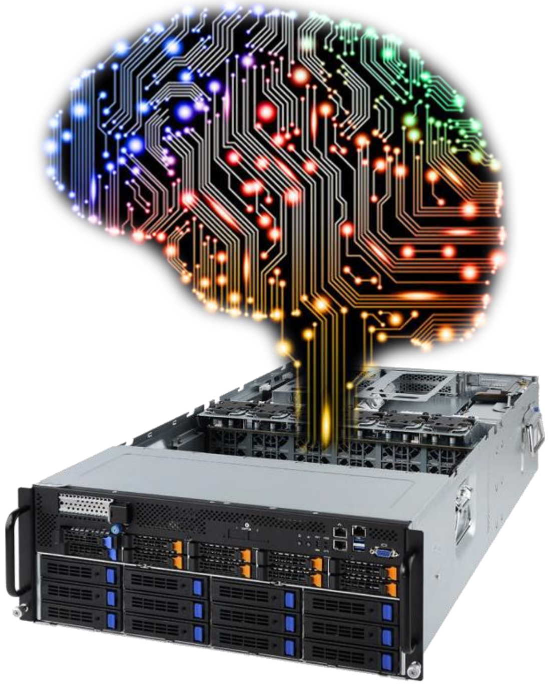

car 98%
car 94%
person 99%
GRYCAN

What's that?

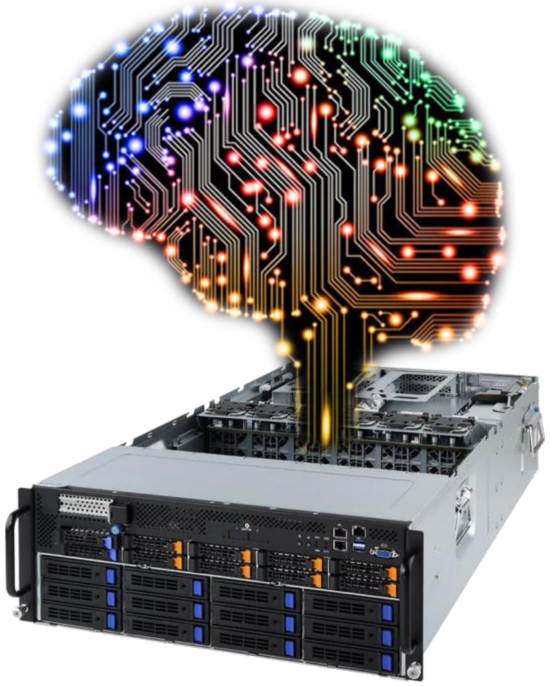




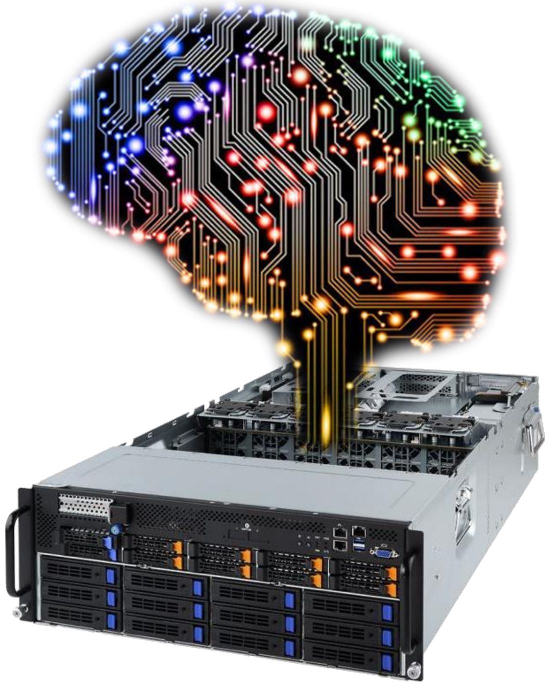
WHAT'S THAT?



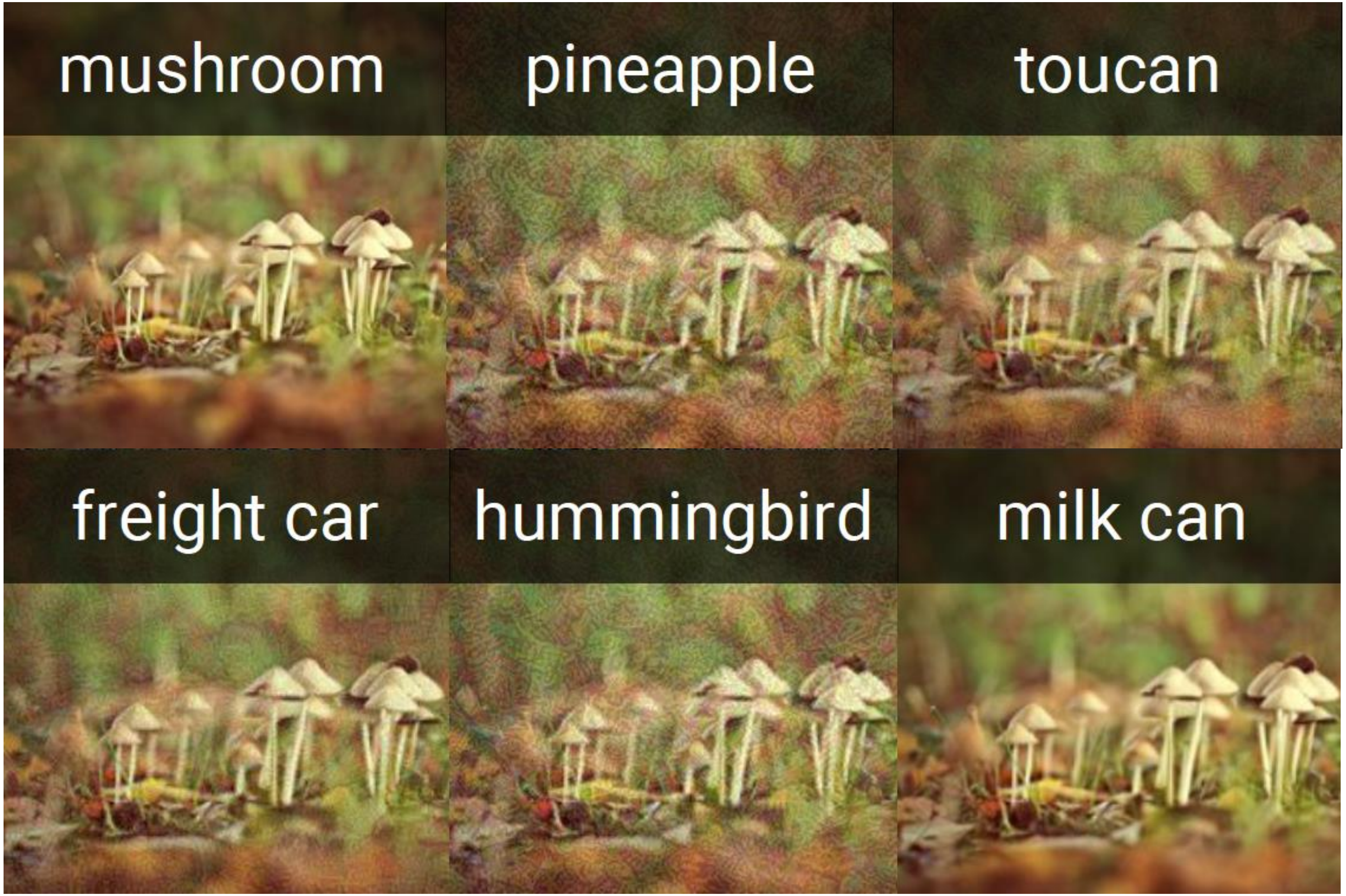
What's that?

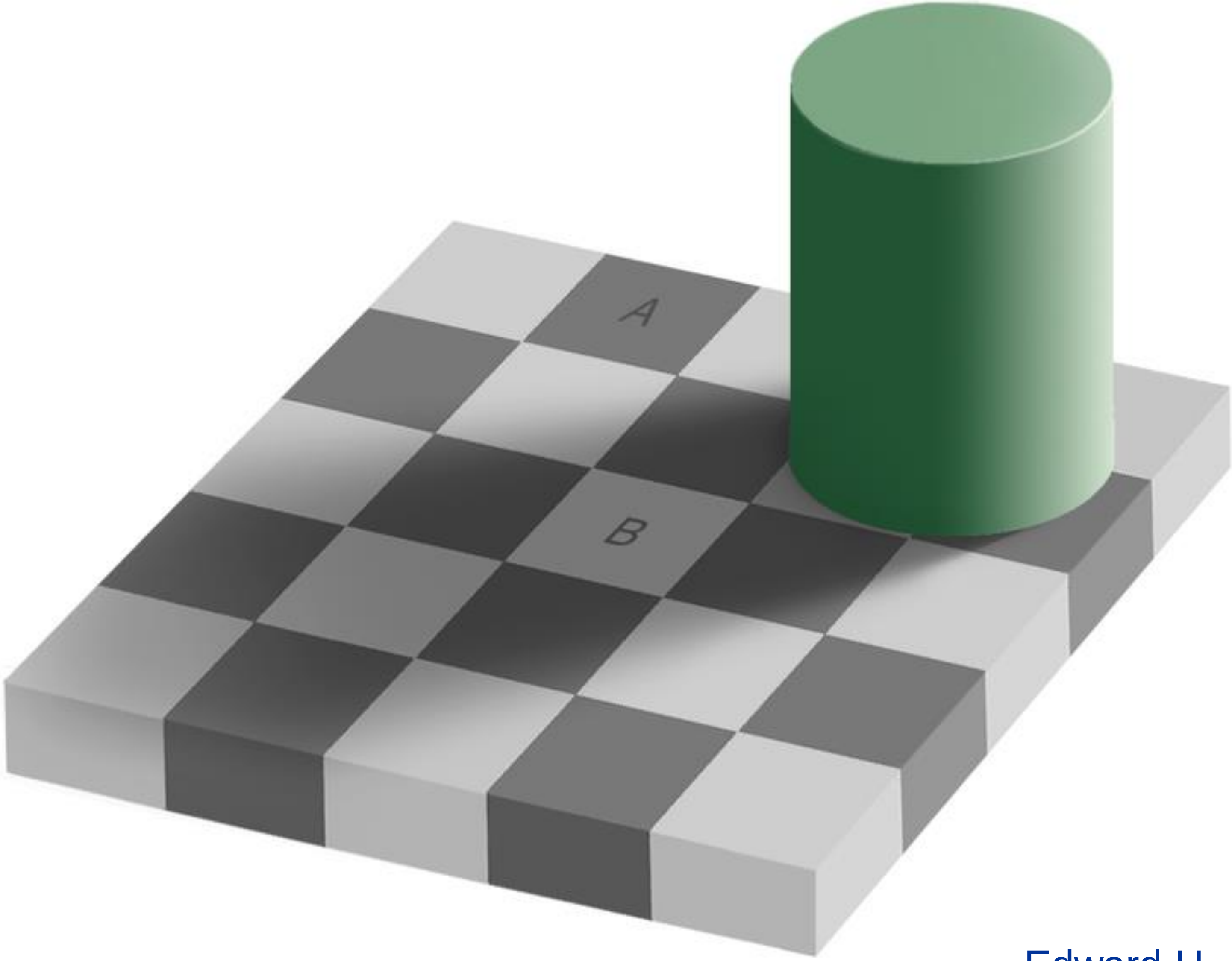


WHAT'S THAT?

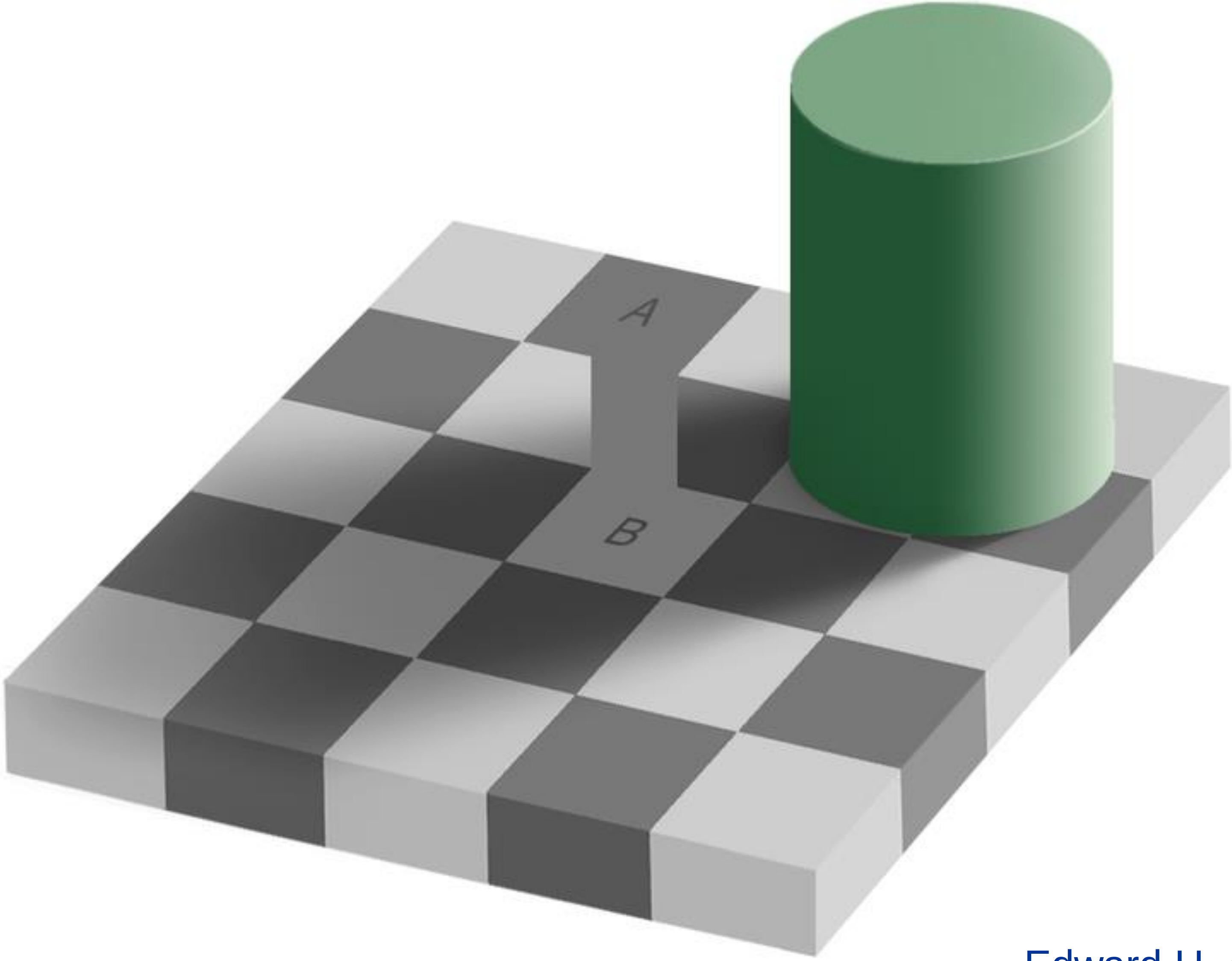


ADVERSARIAL EXAMPLES





Edward H. Adelson



Edward H. Adelson

DUBROVNIK

DUBROVNIK – DEEP DREAM

Know Your Enemy

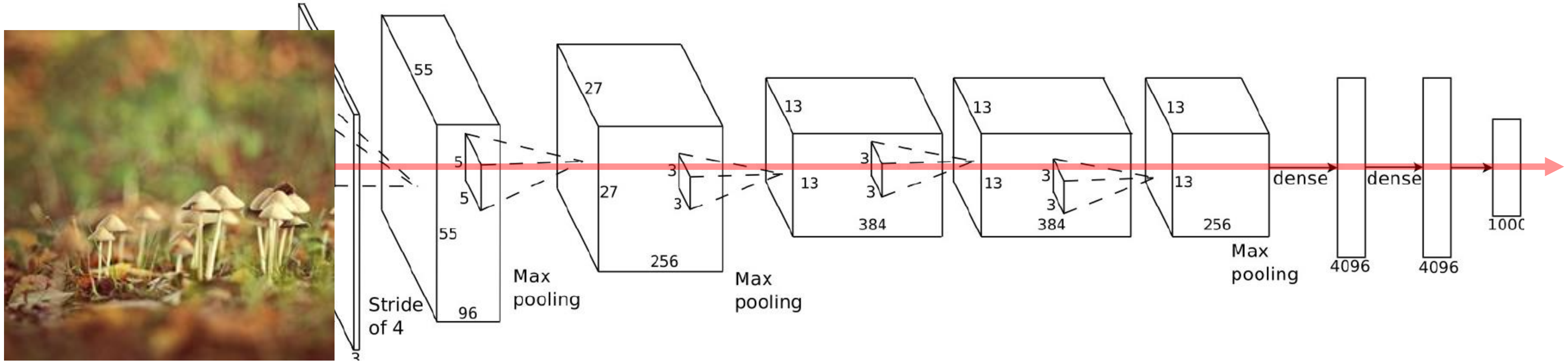
Goal

Knowledge

Capability

ADVERSARY'S GOAL

GENUINE IMAGES



…

Mushrooms

Pineapple

Toucan

…

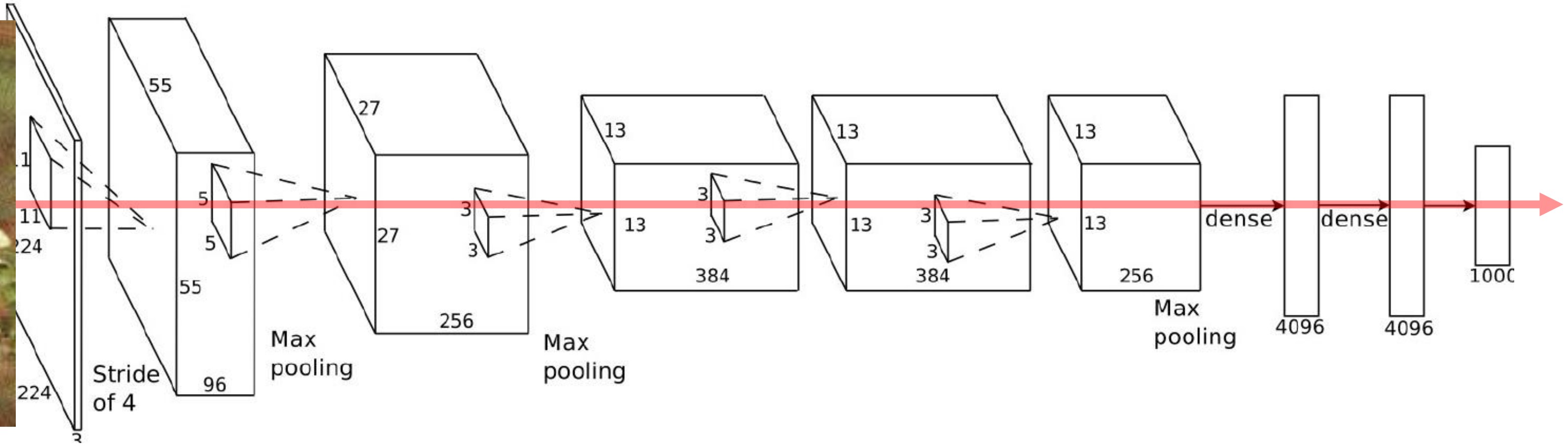
NON-TARGETED ATTACK

+

=

Goal

NON-TARGETED



…

~~Mushrooms~~

<whatever>

…

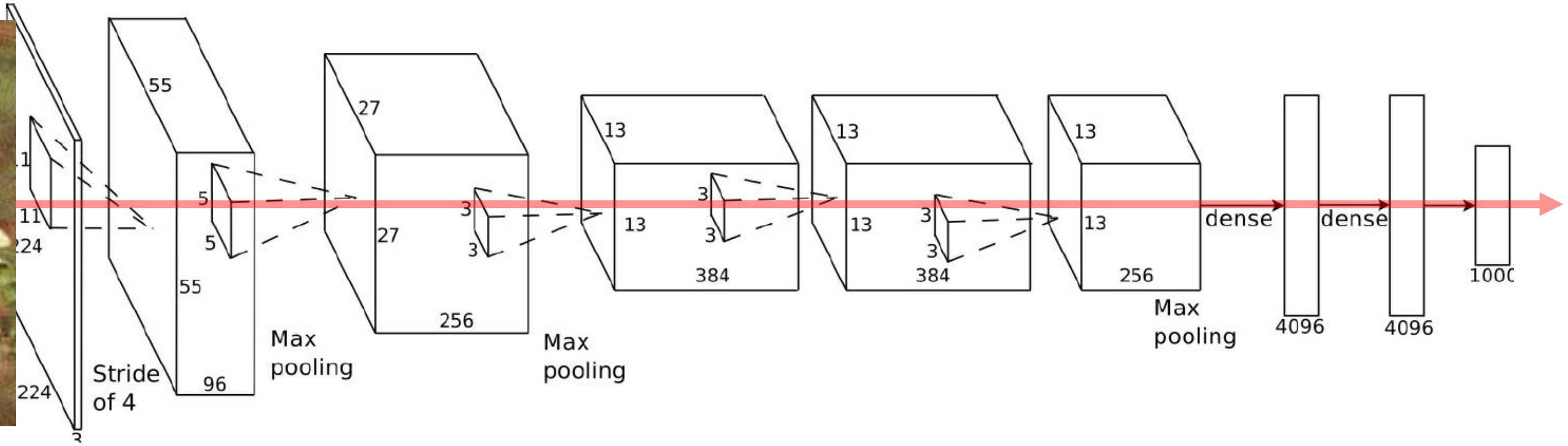
TARGETED ATTACK

+

=

Goal

TARGETED



...

~~**Mushrooms**~~

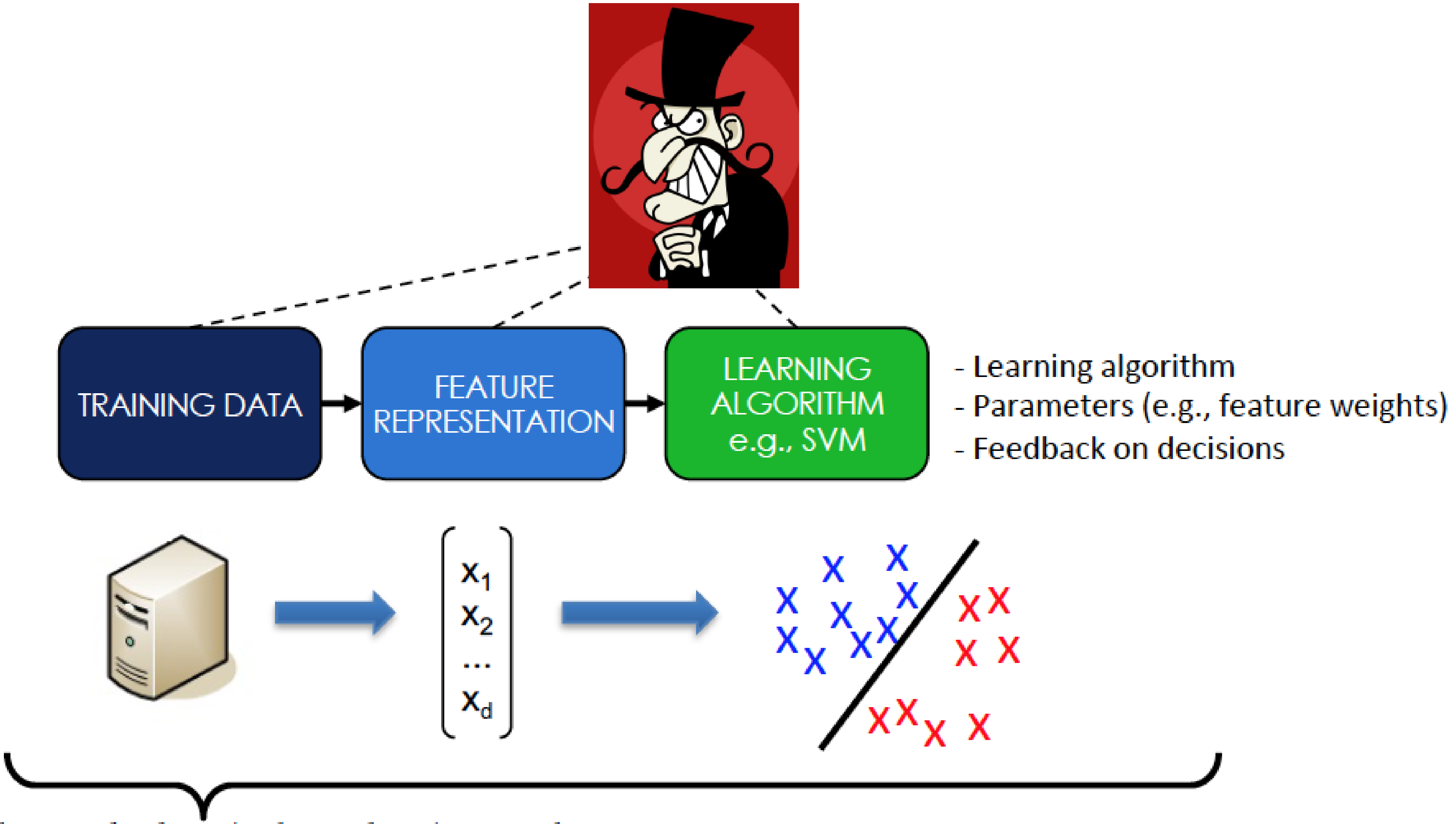
Toucan

...

Goal

Knowledge

Capability



Perfect-knowledge (white-box) attacks

– upper bound on the performance degradation under attack

Slide credit: Biggio

Attacking Deep Neural Networks

- Attacks are possible:
    - if you have the model [1,2]
    - **if you have access to input and output only! [3]**

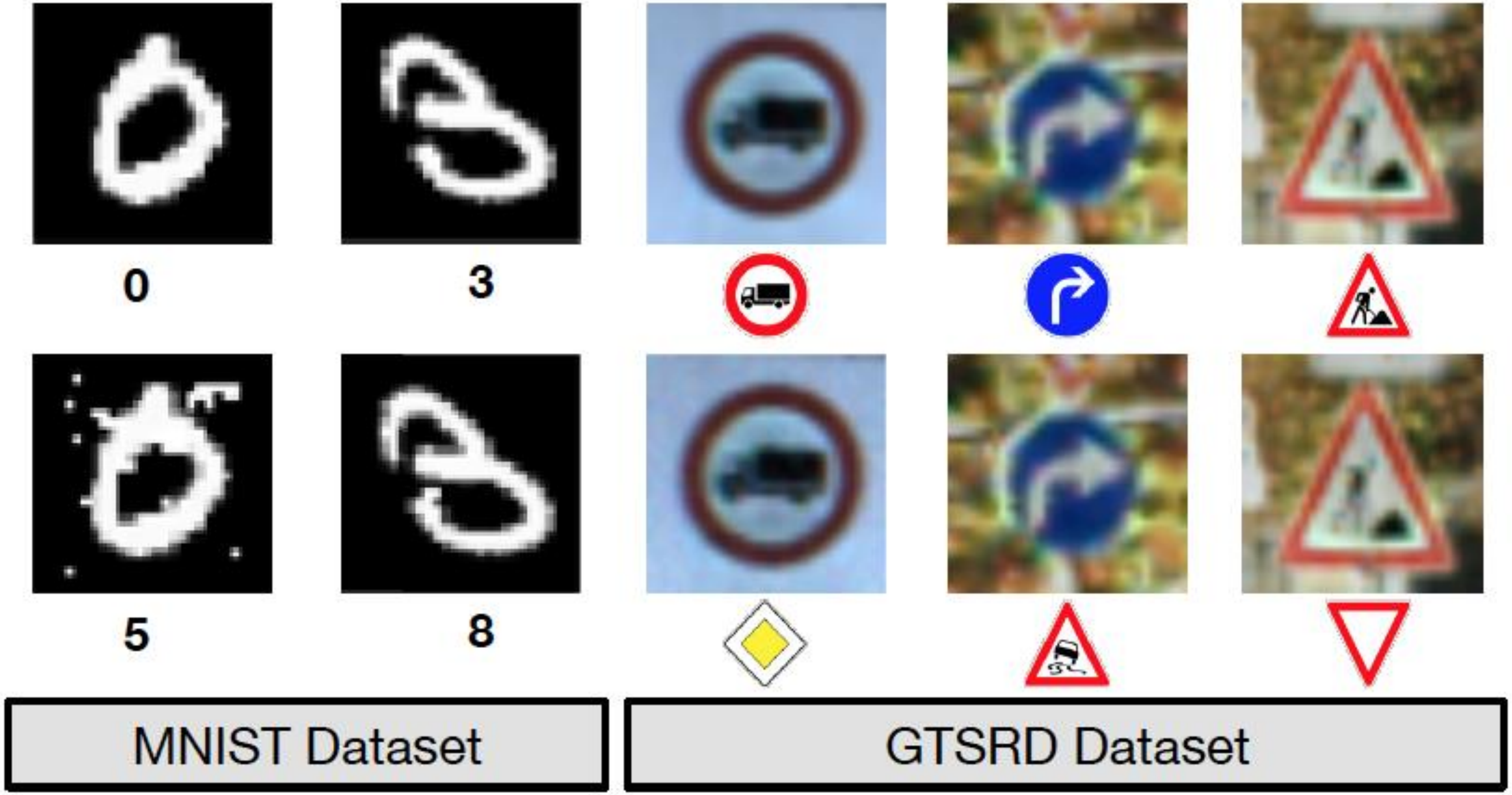
Error Rate: **84.24%** **88.94%** **96.19%**

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*(2013).

[2] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[3] Papernot, Nicolas, et al. "Practical black-box attacks against deep learning systems using adversarial examples." *arXiv preprint arXiv:1602.02697* (2016).

BLACK BOX ADVERSARIAL EXAMPLE ATTACKS



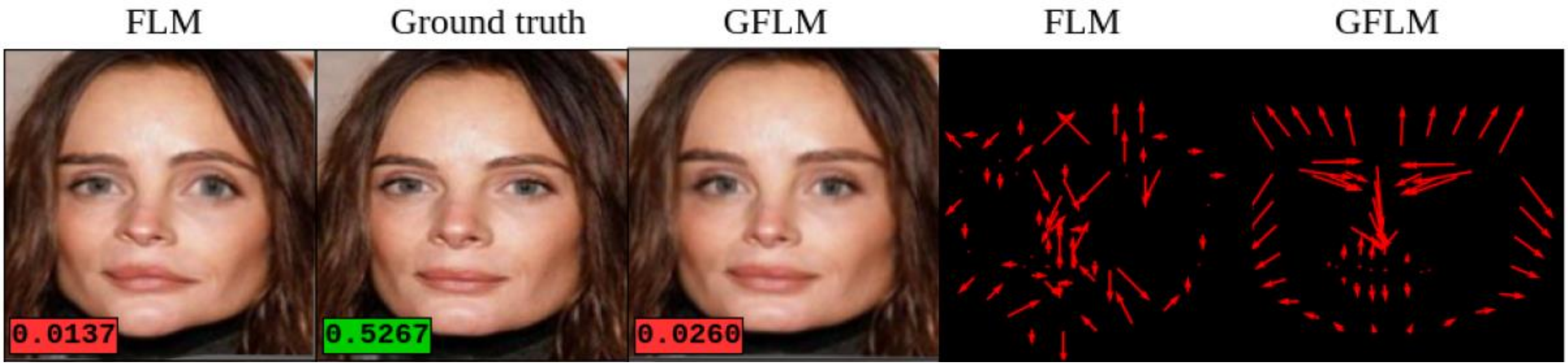
Practical Black-Box Attacks against Machine Learning
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami

Attacking Face Recognition Systems

ADVERIAL FACES

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi

Adversarial Faces



Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi

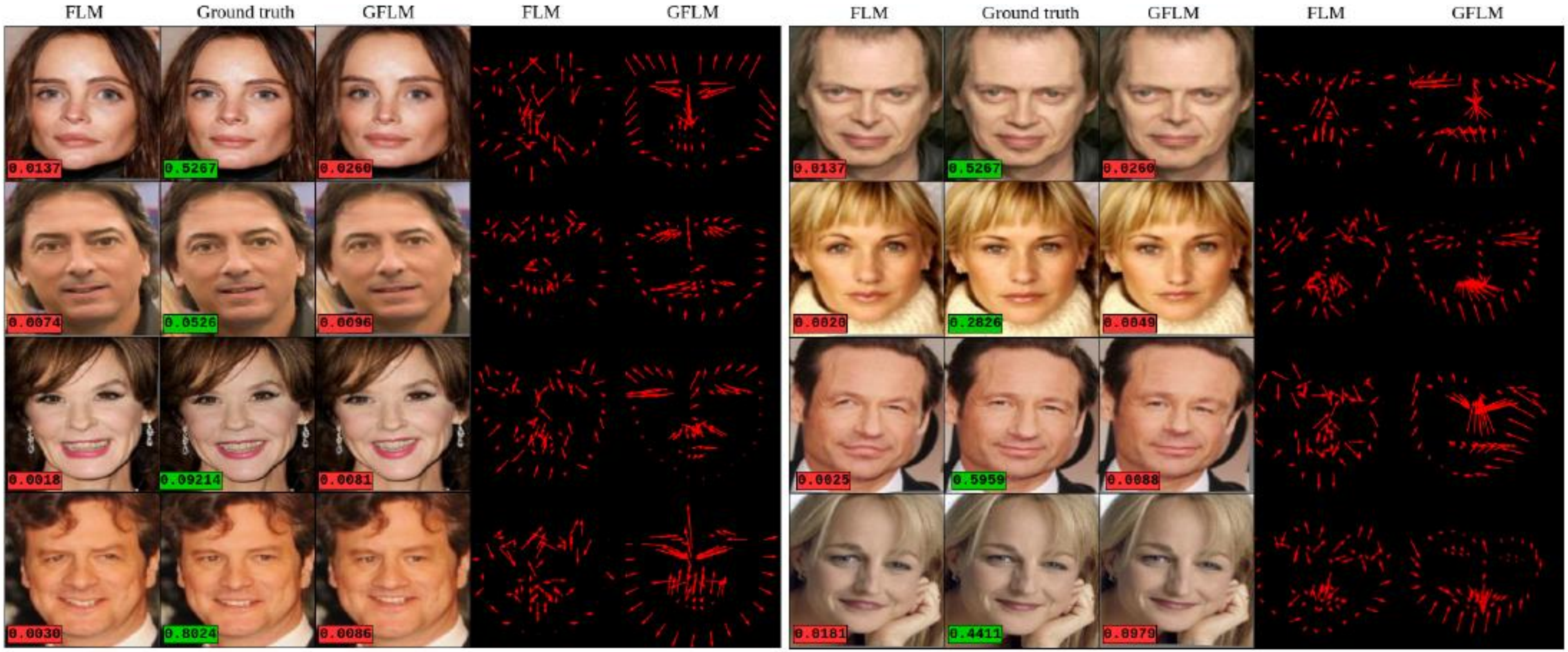
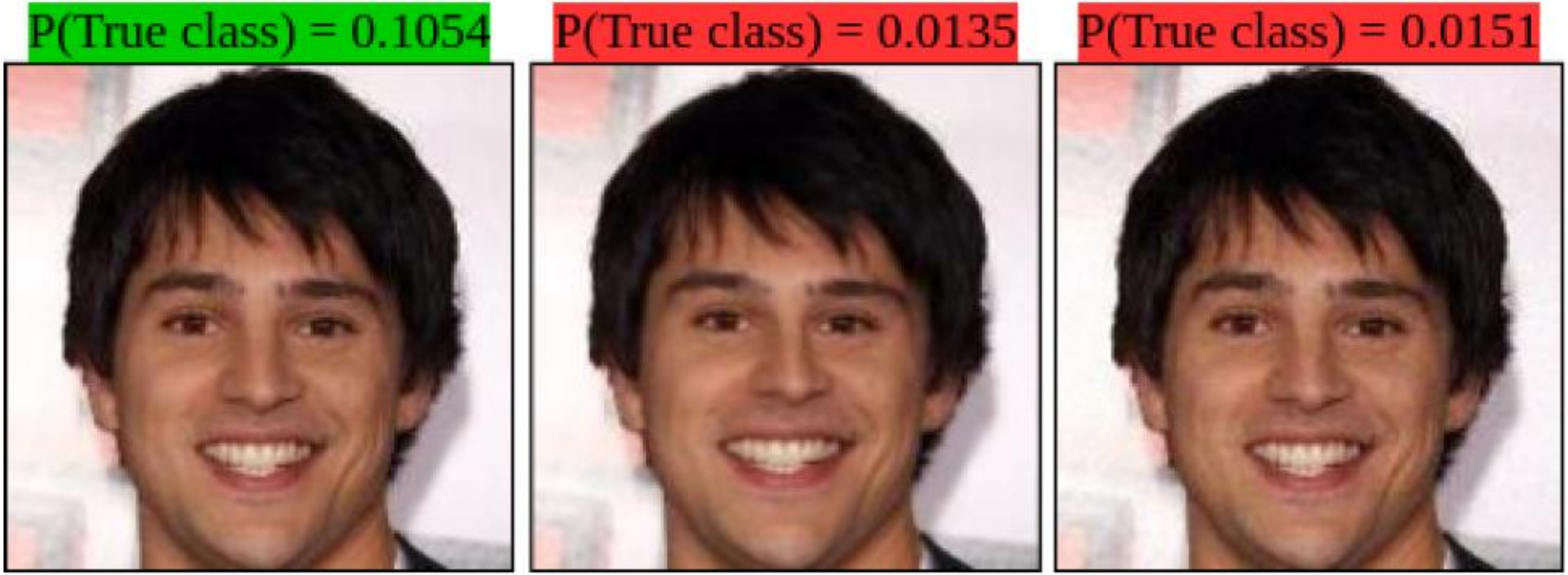


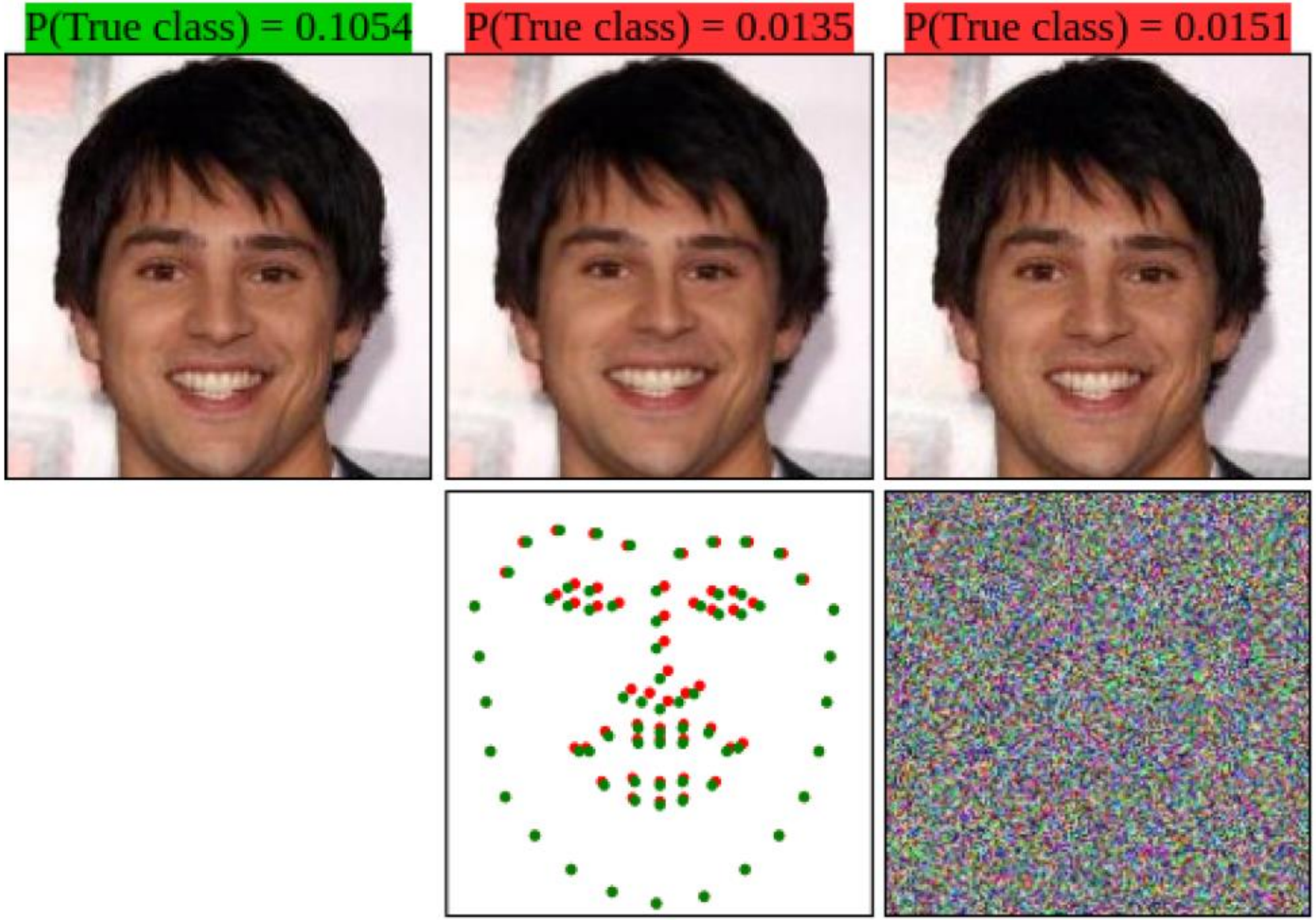
Figure 5. Examples of the adversarial faces generated using FLM and GFLM. For each subject, five images are shown including the original face image (middle face), the result of GFLM (right face), the result of FLM (right image), displacement field f for GFLM (left field) and displacement field f for FLM (right field). Tags on the bottom left of images show the probability of the true class. Green and red tags denote the correct and incorrect classified samples respectively.

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi

Adversarial Faces



Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi



Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi

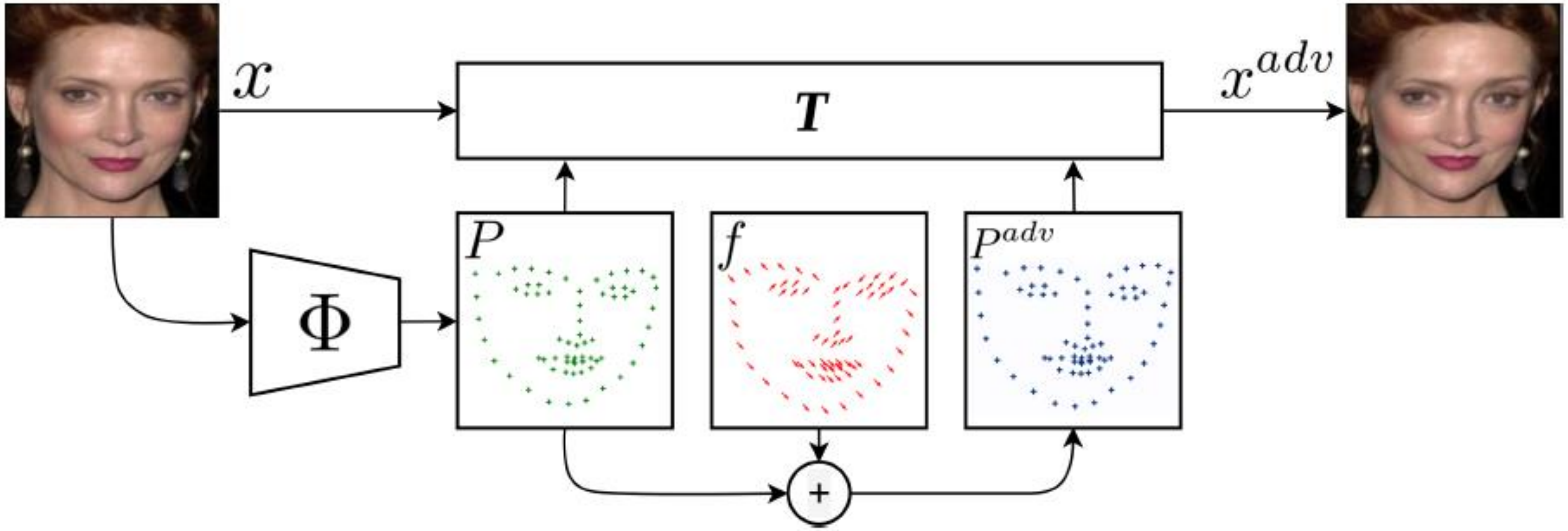
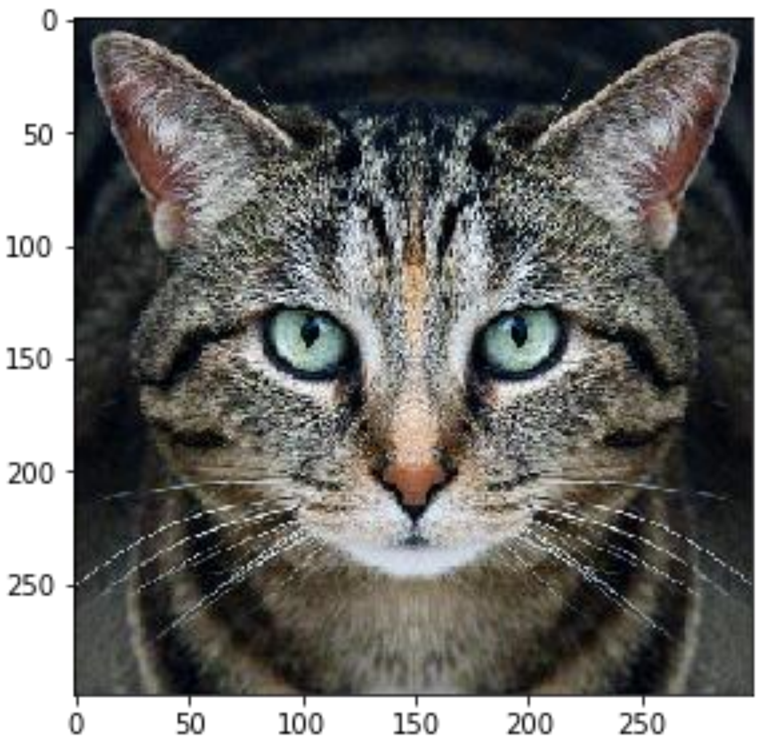


Figure 2. The proposed method optimizes a displacement field f to produce adversarial landmark locations P^{adv}. The spatial transformation T transforms the input sample to the corresponding adversarial image x^{adv} such that $\Phi(x^{adv}) = \Phi(x) + f$, and a state-of-the-art face recognition model g miss-classifies the transformed image x^{adv}.

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi

Attacking in Real World

ADVERSARIAL IMAGE



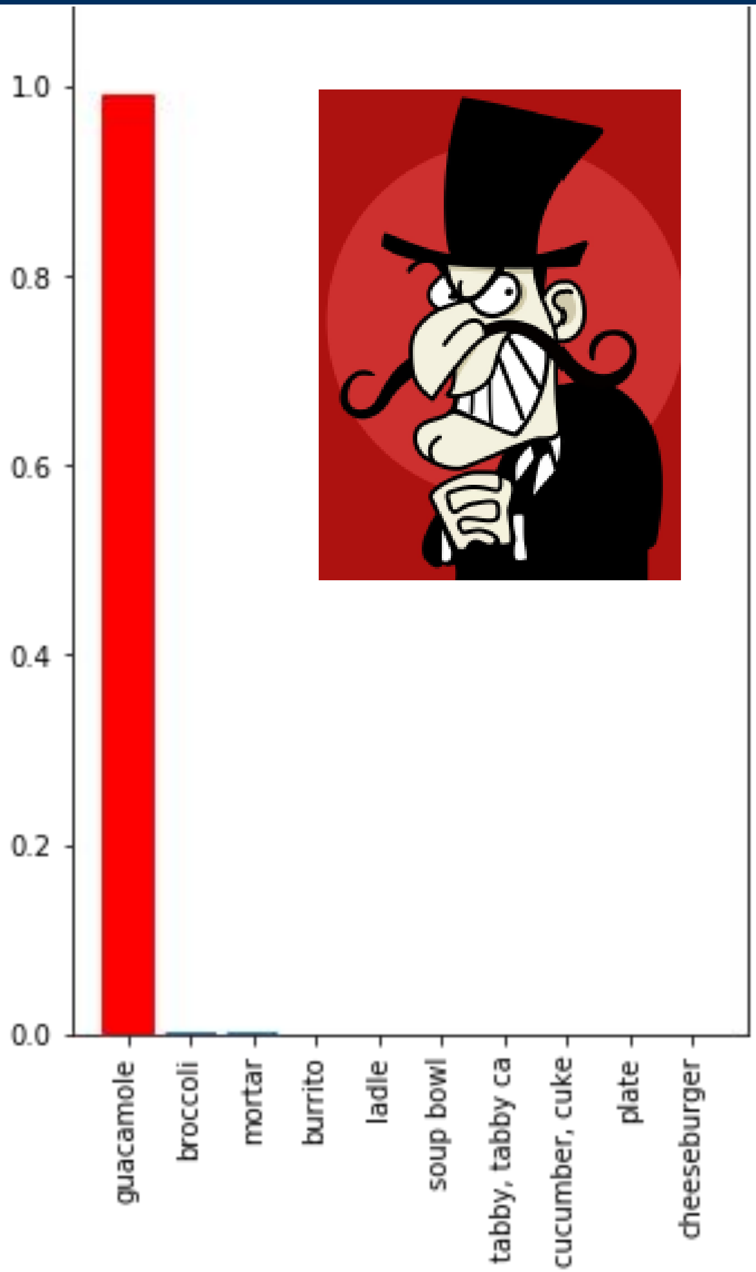
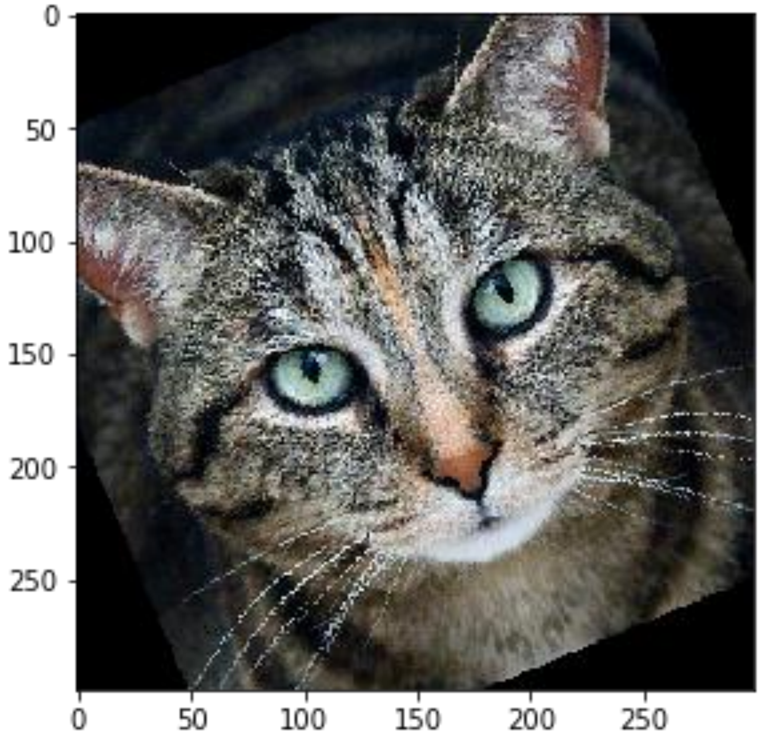


Photo: labsix

Rotate Adversarial Image



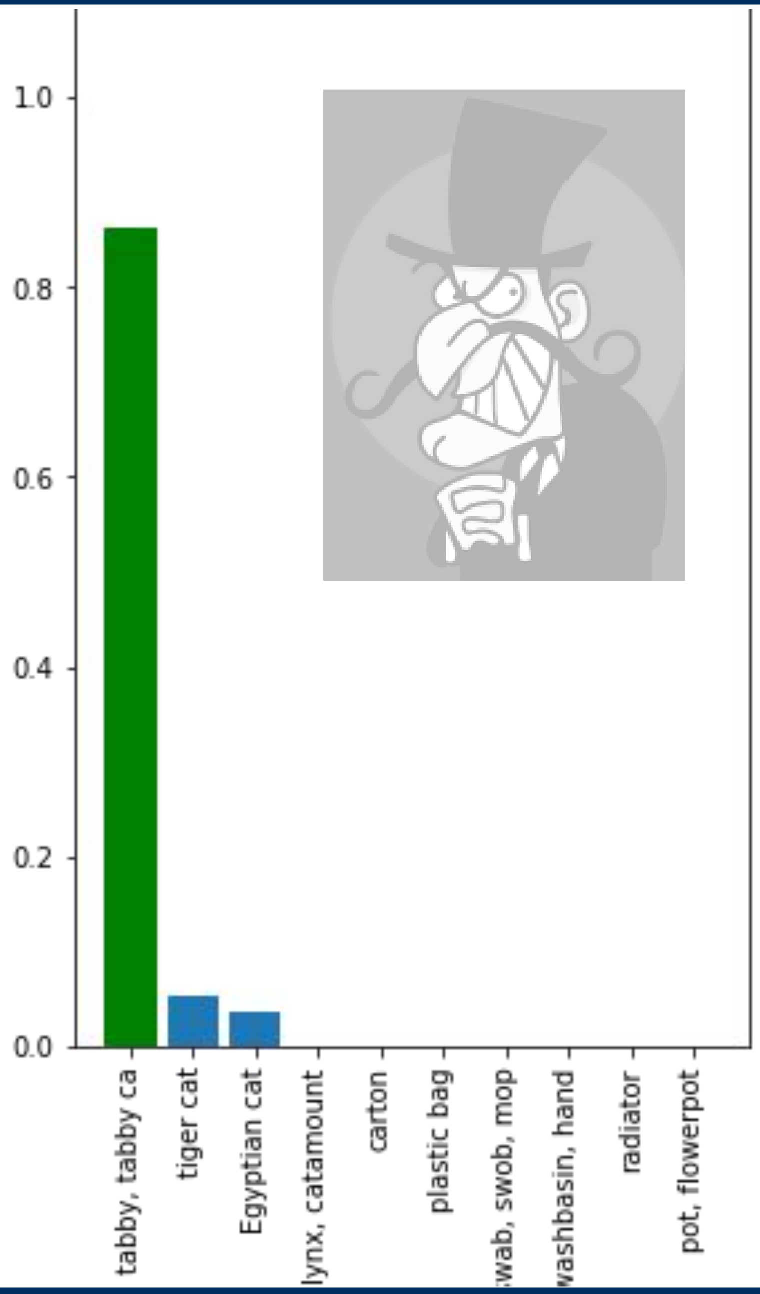
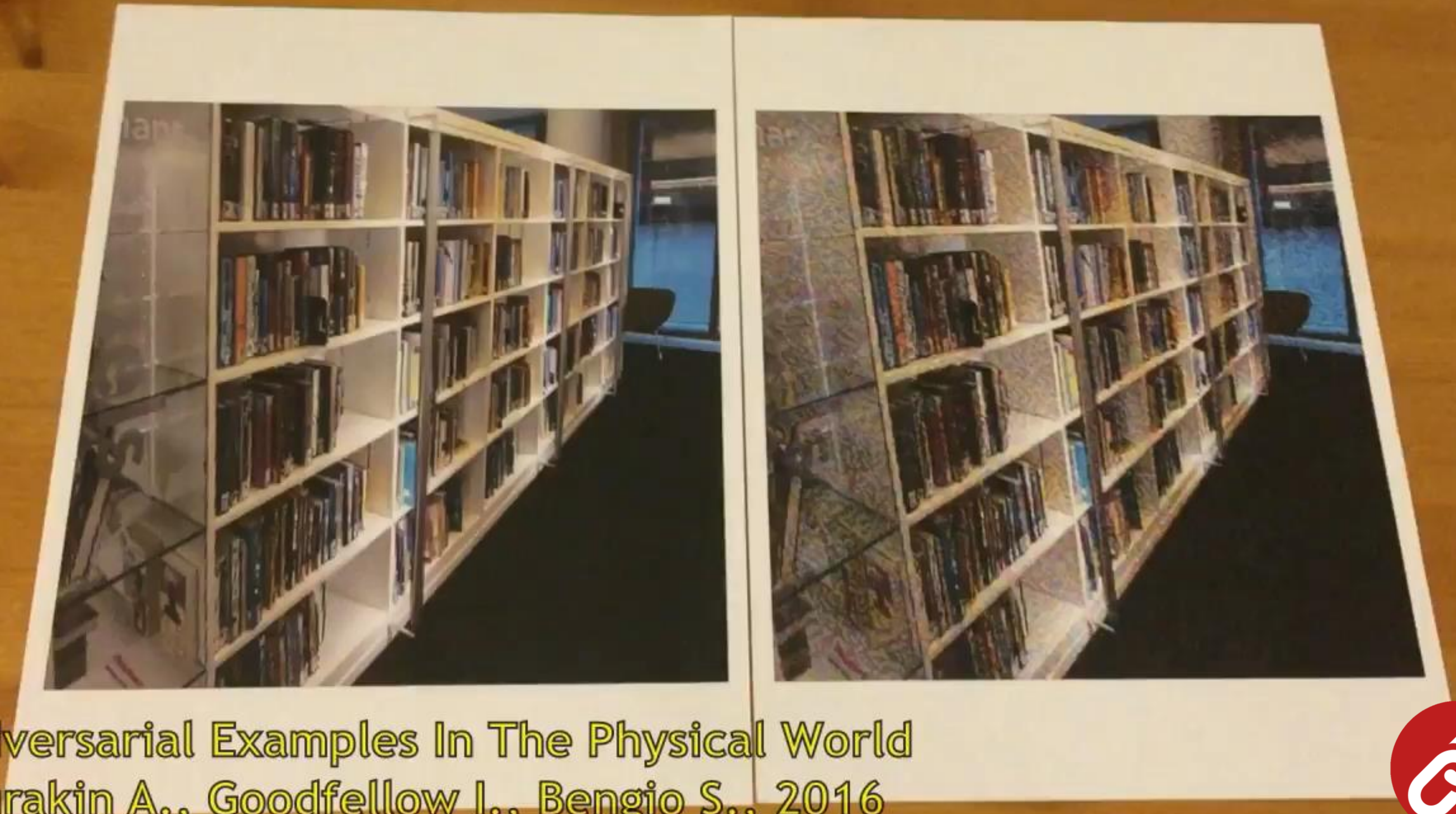
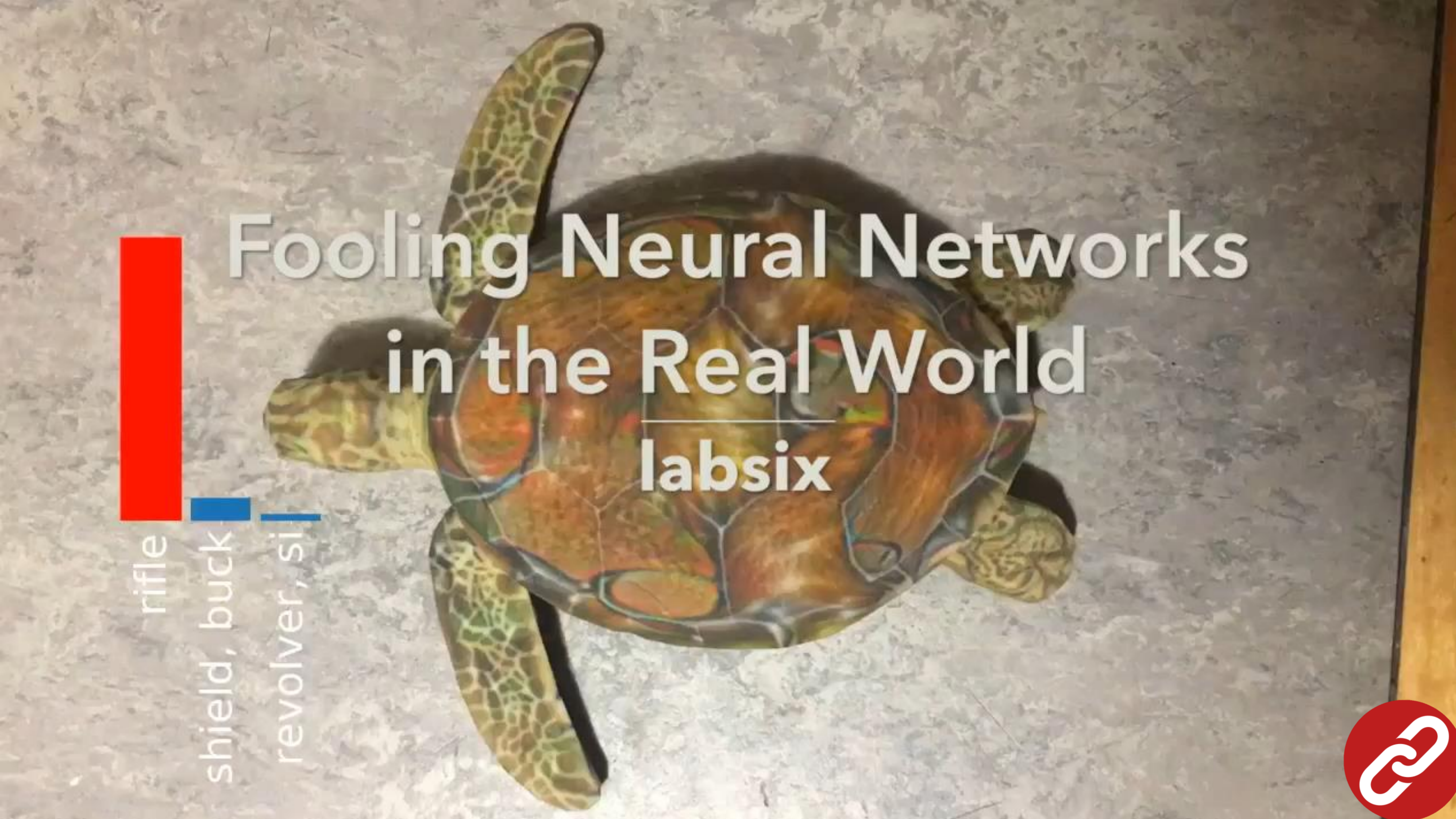


Photo: labsix



Adversarial Examples In The Physical World
Kurakin A., Goodfellow I., Bengio S., 2016


Fooling Neural Networks in the Real World
labsix
rifle
shield, buck
revolver, si

Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)

Robust Physical-World Attacks on Deep Learning Models
Eykholt, Evtimov, Fernandes, Bo Li, Rahmati, Xiao, Prakash, Kohno, Song

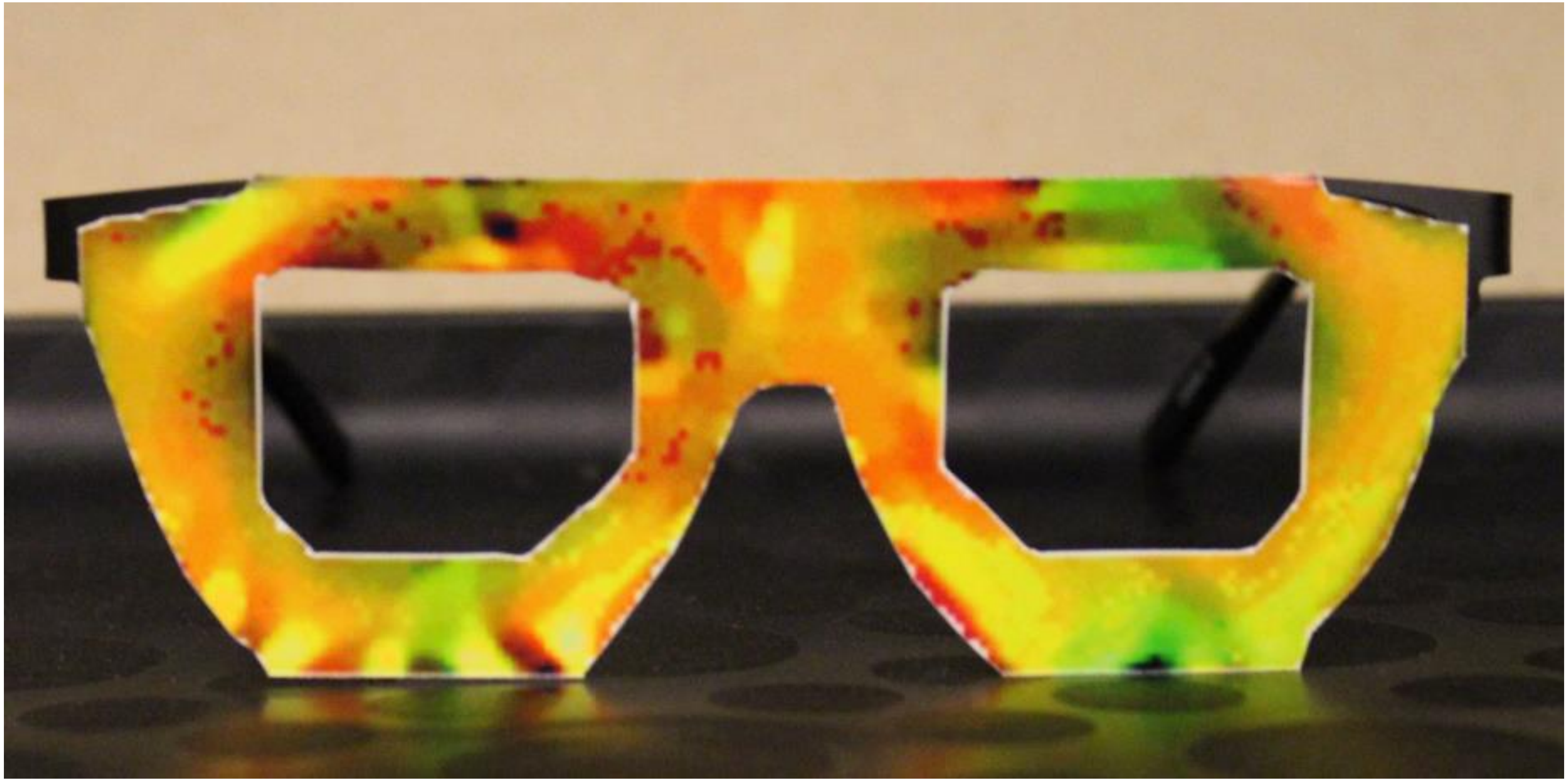
Fig. 4: An example of digital dodging. Left: An image of actor Owen Wilson, correctly classified by VGG143 with probability 1.00. Right: Dodging against VGG143 using AGN's output (probability assigned to the correct class: < 0.01).

Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

ATTACKING DNN IN REAL WORLD

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

ATTACKING DNN IN REAL WORLD



Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

ATTACKING DNN IN REAL WORLD

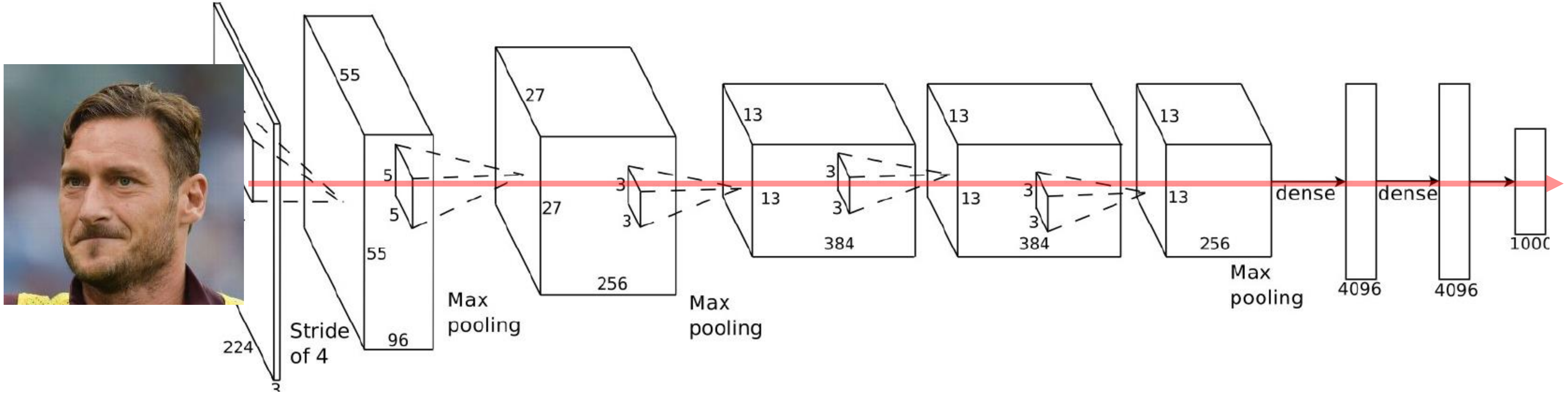
(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

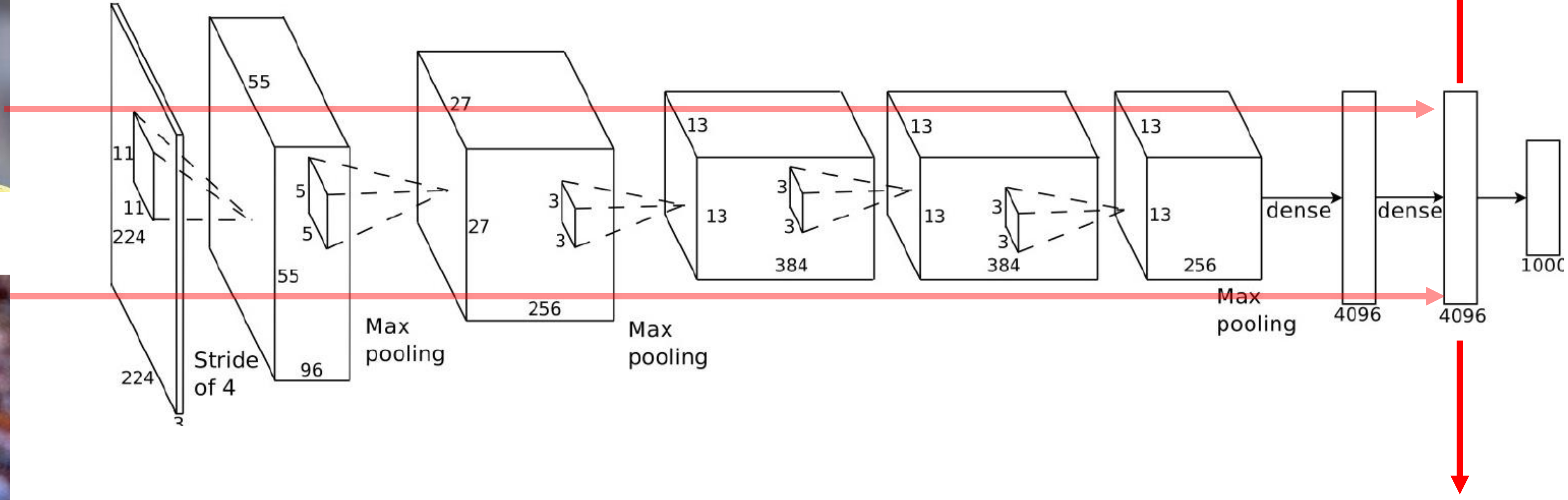
Attacking Face Verification Systems

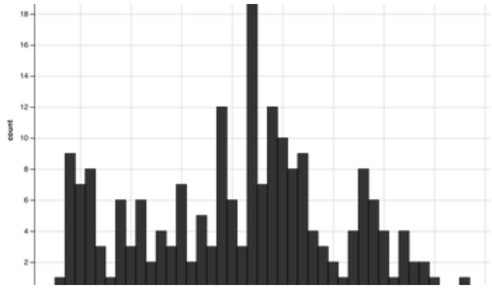
FACE RCOGNITION

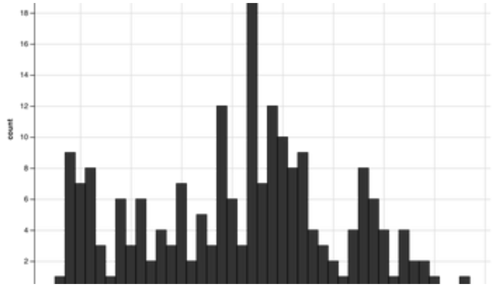


ID1
ID2
ID3
...
ID10
...
IDn

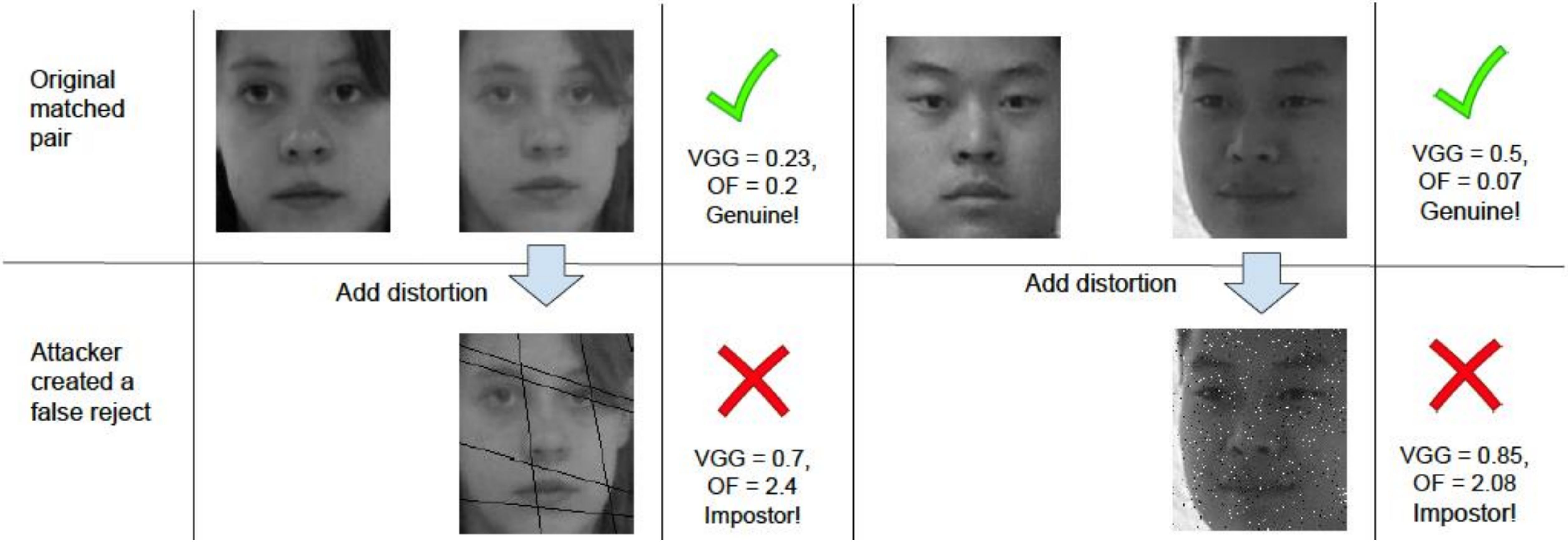
Face Verification



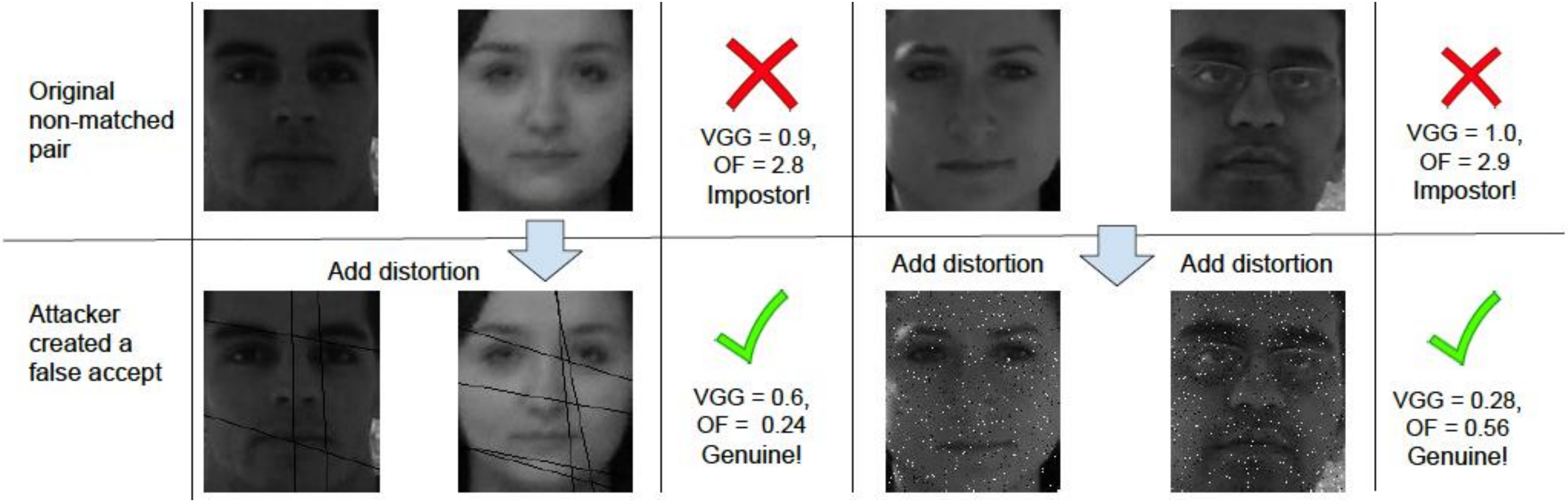


FACE VERIFIATION



Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks
Goswami, Ratha, Agarwal, Singh, Vatsa

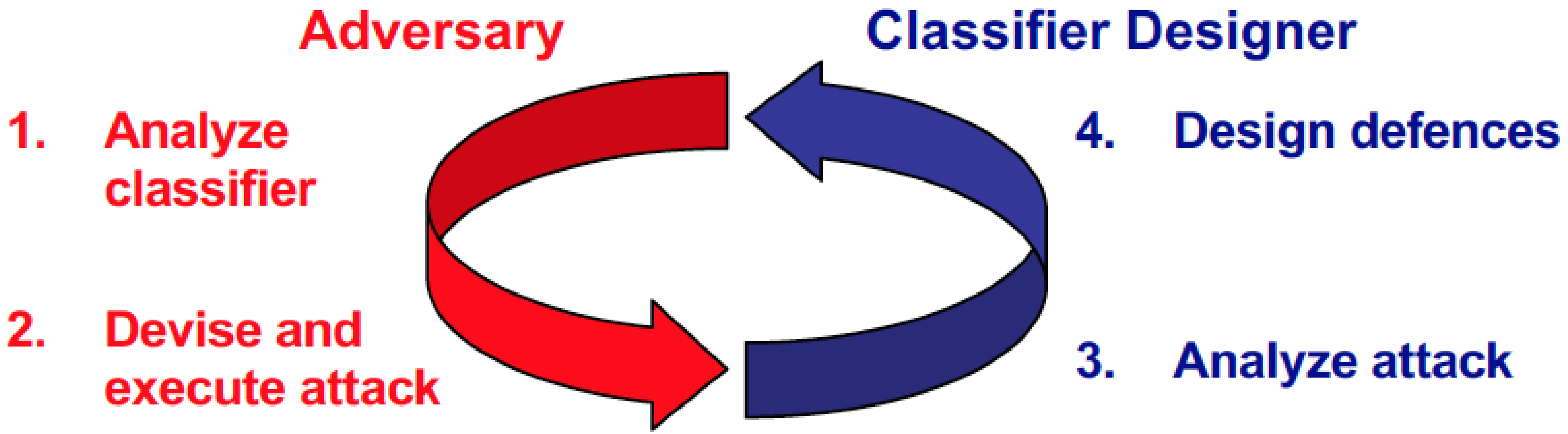
FACE VERIFIATION



Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks
Goswami, Ratha, Agarwal, Singh, Vatsa

Adversary-Aware Machine Learning

ADVERSARY-AWARE MACHINE LEARNING



Machine learning system should be aware of the *arms race* with the adversary

Security evaluation of pattern classifiers under attack
Biggio, Fumera, Roli

NIPS 2018 : Adversarial Vision Challenge (Robust Model Track)

Pitting machine vision models against adversarial attacks.

bethgelab crowdAI Google Brain EPFL Digital Epidemiology Lab

Completed

41429	327	1953
Views	Participants	Submissions

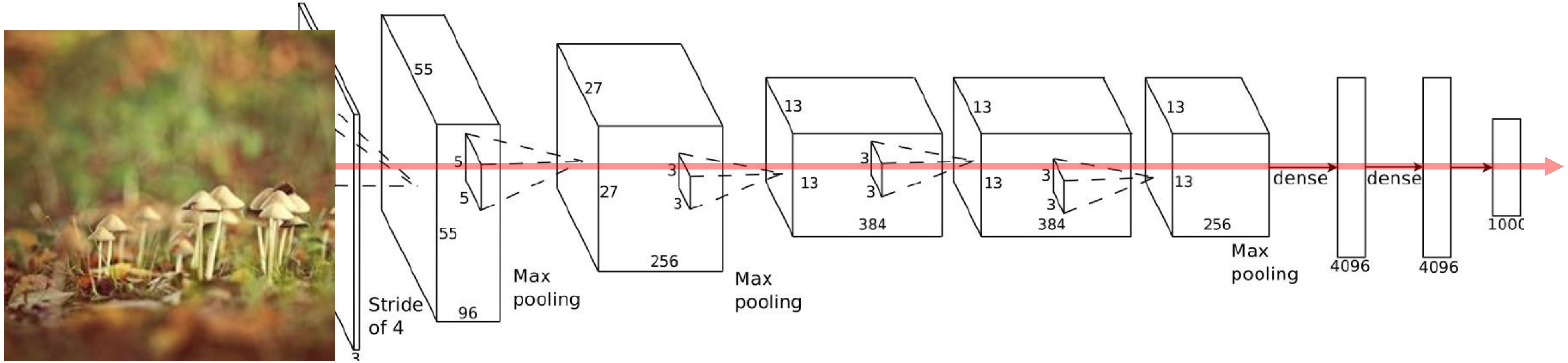
Competition tracks

There will be three tracks in which you and your team can compete:

- Robust Model Track
- Untargeted Attacks Track
- Targeted Attacks Track

Adversarial Example Detection

GENIUNE IMAGES

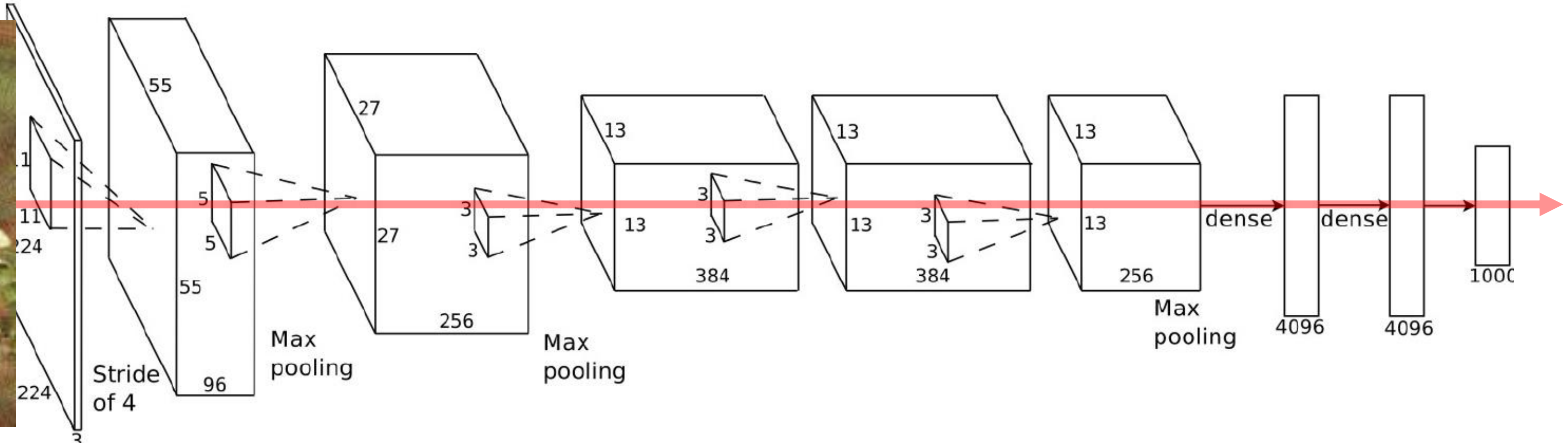


…
Mushrooms
Pineapple
Toucan
…

NON-TARGETED ATTACK

+

=



...

~~Mushrooms~~

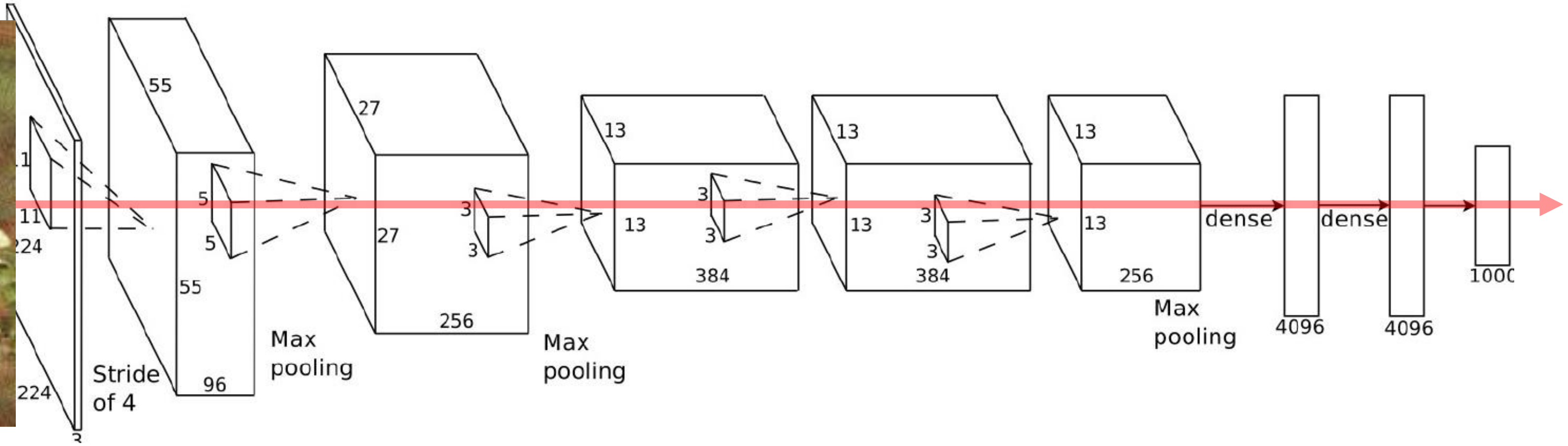
<whatever>

...

DEFENSE

+

=



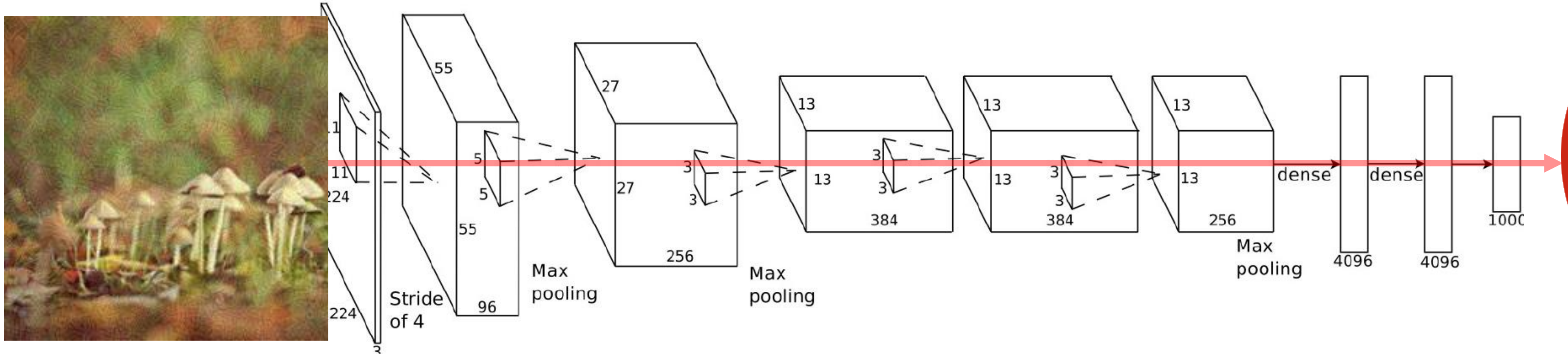
…
Mushrooms
…

Increase robustness

DETECTION

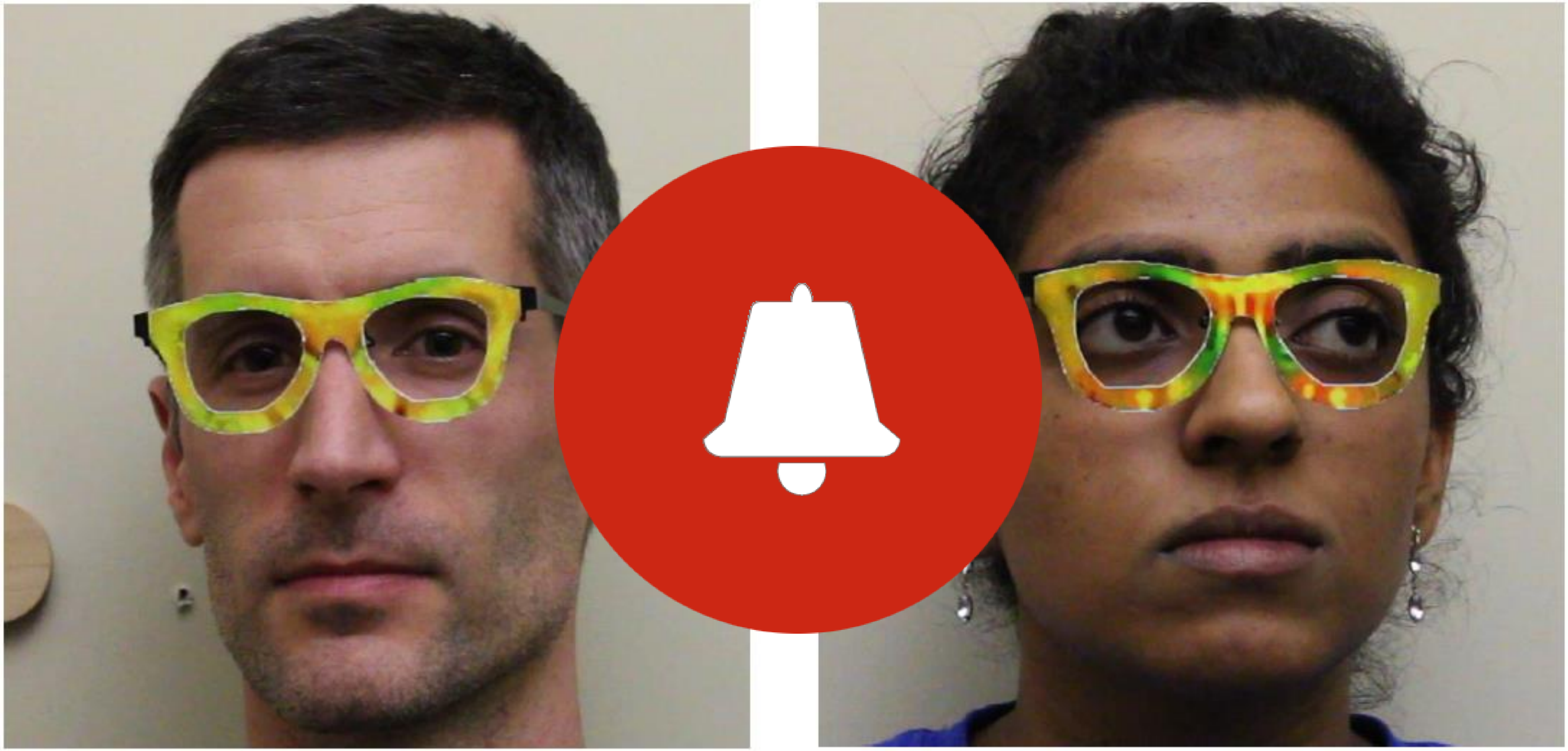
+

=



Attack detection

Adversarial Examples Detection



Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

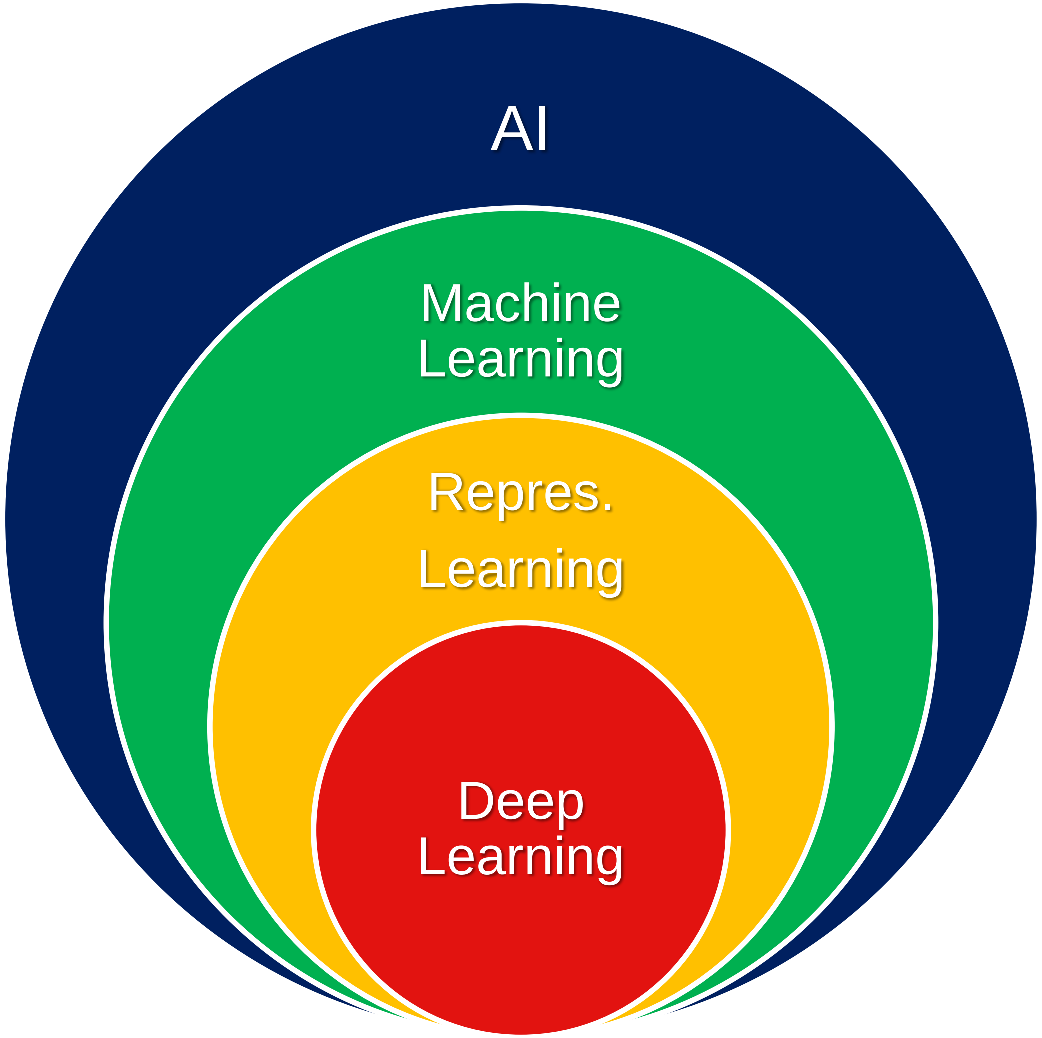
Our Approach

DEEP LEARNING (FROM NATURE)

nature International weekly journal of science

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

Yann LeCun, Yoshua Bengio & Geoffrey Hinton



DEEP LEARNING (FROM NATURE)

nature International weekly journal of science

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

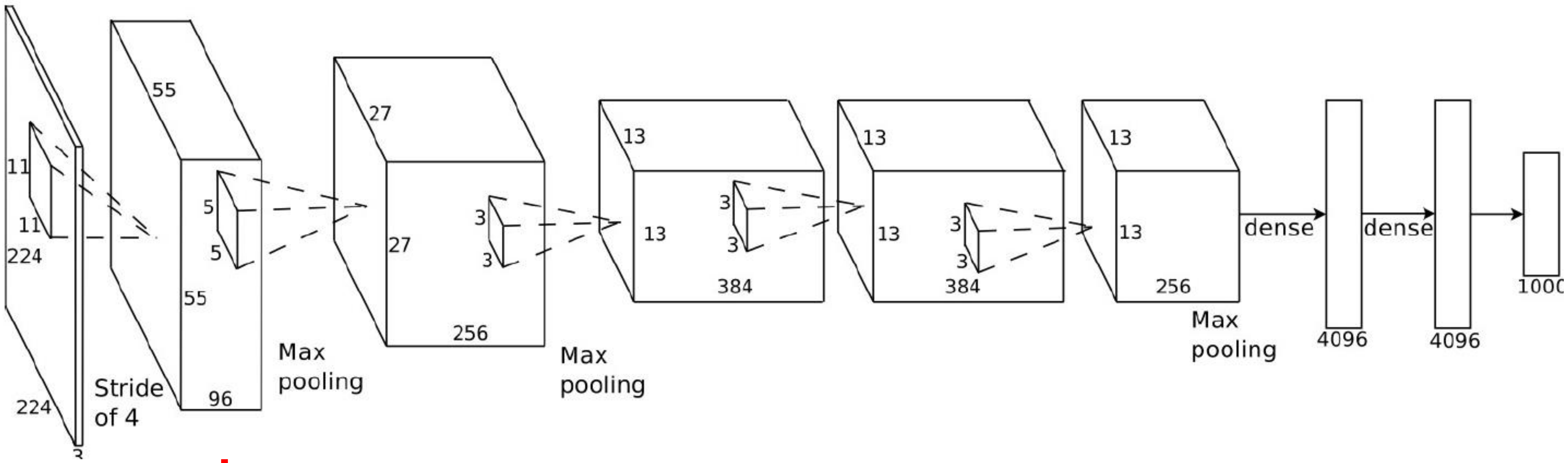
Representation learning methods that

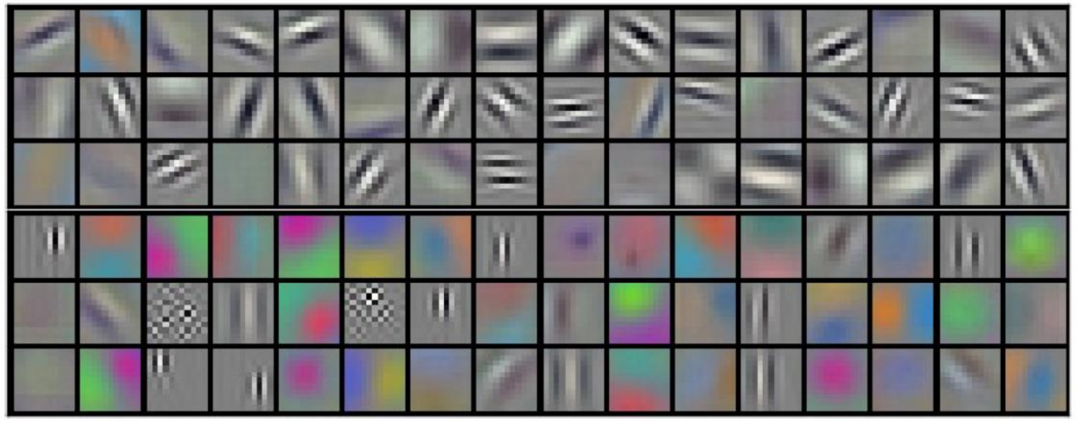
allow a machine to be fed with raw data and to automatically discover the representations needed for detection or classification.

Deep-learning are representation learning methods

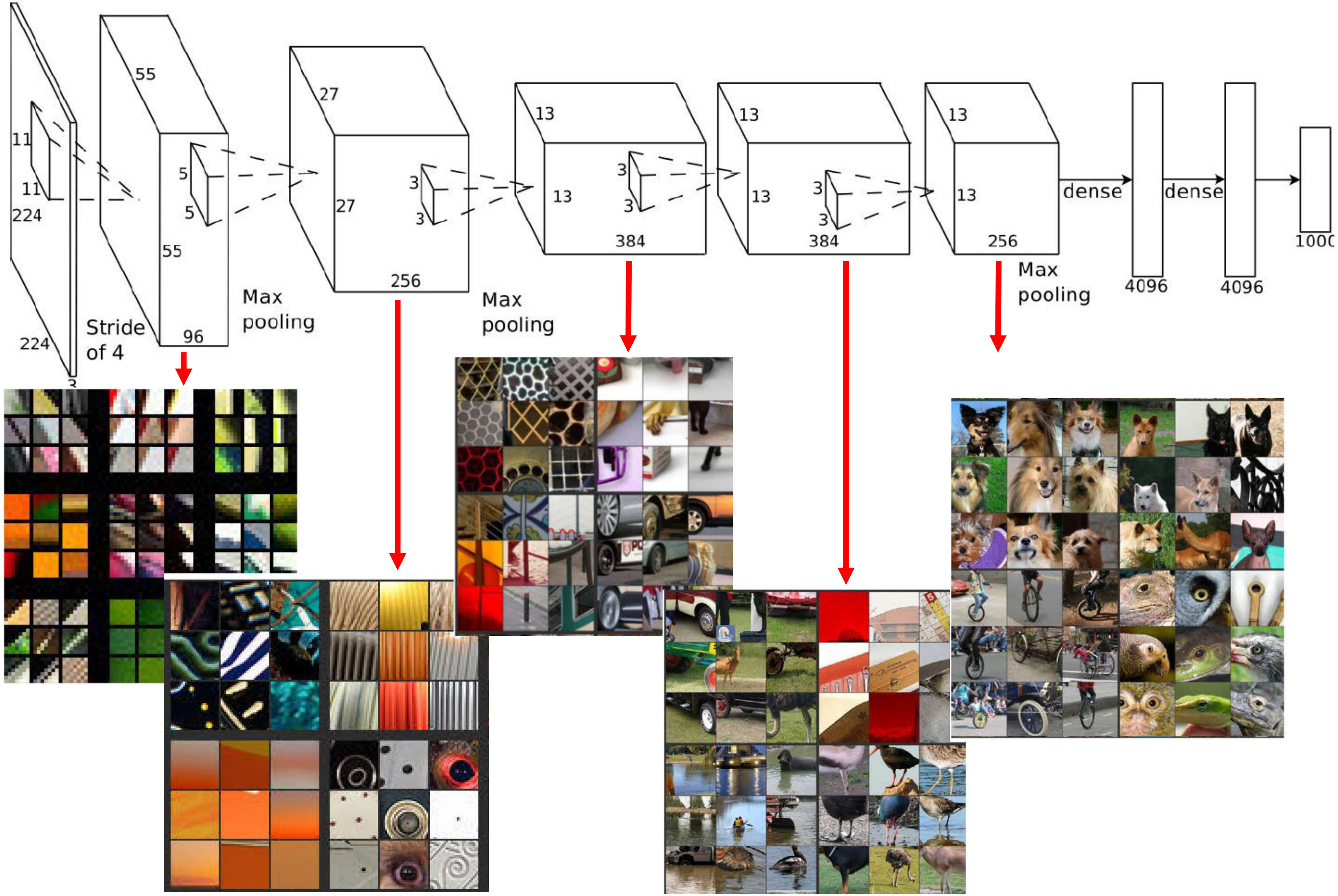
- with **multiple levels** of representation, obtained by
- composing simple but **non-linear modules** that each
- transform the representation at one level
  into a representation at a higher,
  slightly more abstract level.

Multiple Levels Of Abstraction





Multiple Levels Of Abstraction

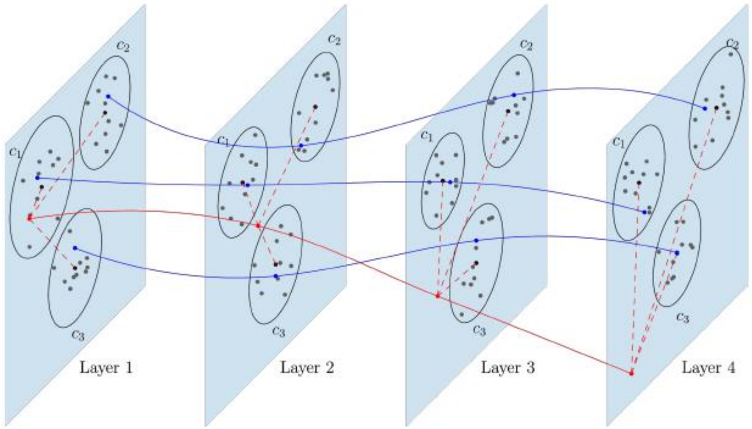


OUR APPROACH

A **detection scheme** for adversarial images based on internal representation (aka *deep features*) of the neural network classifier.

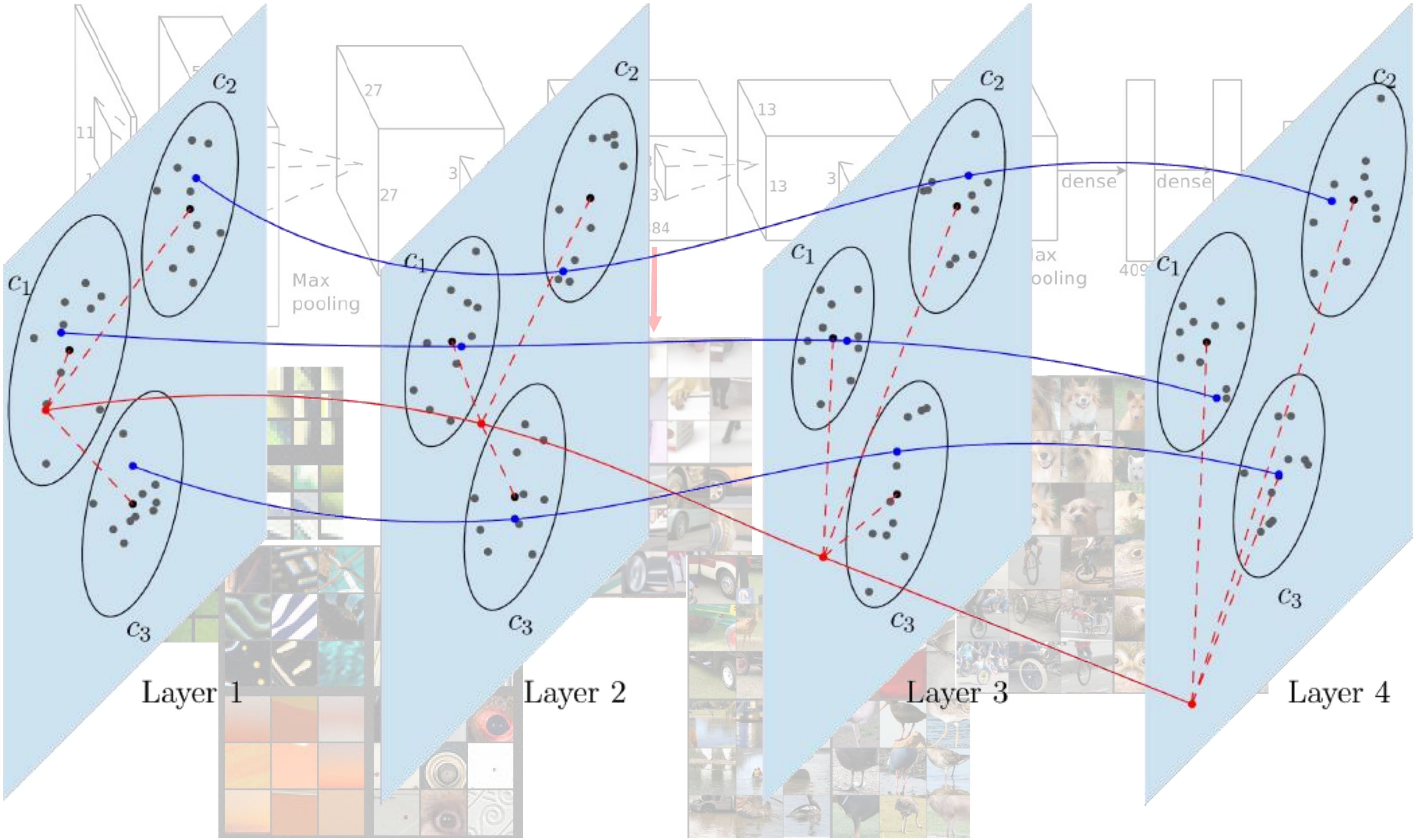
- **Main intuition**: look at the evolution of features, i.e. the path formed by their positions in the feature spaces, during the forward pass of the network.

- **Claim**: The trajectories traced by authentic inputs and adversarial examples differ and can be used to discern them.



Adversarial examples detection in features distance spaces
F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, G. Amato
ECCV WOCM Workshop 2018

Multiple Levels Of Abstraction

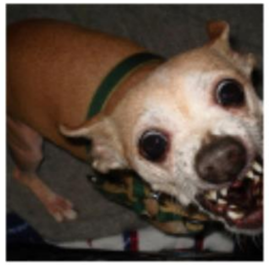


Our Approach: Results

Attacked Model	ResNet-50 pretrained on ILSVRC'12
Crafting Algorithms	L-BFGS, FGSM, BIM, PGD, MI-FGSM
Emb. Pivots	1000 Class (C)entroids / <u>(M)edoids</u> from ILSVRC train set
Emb. Distance Function	L2 / <u>cosine similarity</u> (cos)
Emb. Size	16-length 1000-dim sequences, TRAIN / VAL / TEST = 12k / 1k / 3k
Detector	MLP (2-layer, 100 and 1 neurons) / <u>LSTM</u> (100-dim)
Threat Model	zero-knowledge (attacker not aware of detector)

Method	L-BFGS	FGSM	BIM	PGD	MI-FGSM	macro-AUC
LSTM + M + cos	**.854**	**.996**	.997	.997	.997	**.968**
LSTM + M + L2	.743	**.996**	**.998**	.998	**1.000**	.947
MLP + M + cos	.551	.992	.996	.995	.998	.907
MLP + M + L2	.681	.976	**.998**	**.999**	**1.000**	.931
LSTM + C + cos	.709	.811	.784	.784	.930	.804
LSTM + C + L2	.482	.854	.819	.816	.872	.769
MLP + C + cos	.388	.694	.881	.878	.962	.761
MLP + C + L2	.626	.820	.990	.989	**1.000**	.885

Easy to Identify Adversarial images

Adversarial Image	Generation Algorithm	Actual Class	Fooled Class	Nearest Neighbor	kNN score
	L-BFGS	bikini, two-piece	pomegranate		0.01
	FGS	brassiere, bra, bandeau	Chihuahua		0.01
	FGS	revolver, six-gun, six-shooter	mousetrap		0.00
	L-BFGS	assault rifle, assault gun	Border terrier		0.00

Hard to Identify Adversarial Images

Adversarial Image	Generation Algorithm	Actual Class	Fooled Class	Nearest Neighbor	kNN score
	FGS	chime, bell, gong	barometer		**0.13**
	L-BFGS	basenji	Arctic fox, white fox, Alopex lagopus		**0.13**
	FGS	Greater Swiss Mountain dog	Bernese mountain dog		**0.11**
	FGS	jeep, landrover	pickup, pickup truck		**0.11**

Other Detection Approaches

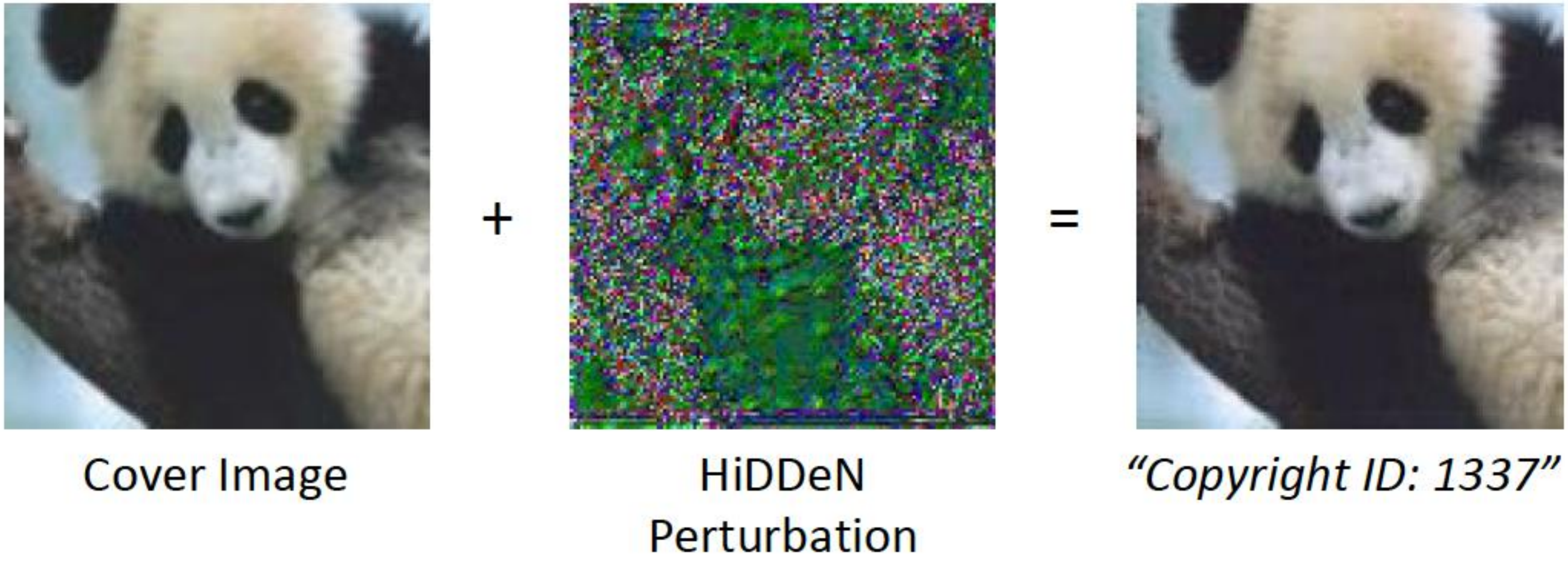
- ***Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*** *[2017]*
  Nicholas Carlini, David Wagner
- ***On Detecting Adversarial Perturbations*** *[2017]*
  Jan Hendrik Metzen, Tim Genewein, Volker Fischer, Bastian Bischoff
- ***Trace and detect adversarial attacks on CNNs using feature response maps*** *[2018]*
  Mohammadreza, Friedhelm, Thilo
- ***Adversarial examples detection in features distance spaces*** *[2018]*
  F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, G. Amato

Related Topics

Detecting Face Morphing Attacks

Detection of Face Morphing Attacks by Deep Learning
C. Seibold, W. Samek, A. Hilsmann, P. Eisert

ADVERSARIAL EXAMPLES DETECTION



HiDDeN: Hiding Data With Deep Networks
Jiren Zhu, Russell Kaplan, Justin Johnson, Li Fei-Fei

THANKS!

Questions are welcomed

Fabrizio Falchi
fabrizio.falchi@cnr.it

Conclusions

- Machine Learning and Deep Learning in particular can be attacked
  - Slightly modifying images but also in real world
  - Even if our neural network is a black box for the enemy
- Many approaches have been proposed to make DL more **robust**
- Adversarial examples **detection** is its early stages
- We need **adversary-aware machine learning**