Malicious Facial Image Processing and Counter-measures: A review

Jean-Luc DUGELAY
jld@eurecom.fr

EURECOM
Sophia Antipolis

1

Motivations

Spoofing is when someone or something pretends to be something else…

- to usurp an identity (or to evade)
  - **biometrics**, video surveillance
- to harm someone
  - fake news
- to discredit a system
  - adversarial attacks
- to reveal an identity (de-anonymization)
  - video surveillance, media

2

Why Facial is one of the most popular biometric trait?

- Human compliant
- ICAO[1] compliant
- No user interaction is required
- Enhanced security and high accuracy
- Can extract other characteristics (age, gender, emotion etc.)

[1] International Civil Aviation Organisation

3

What's New about Facial Recognition Systems?

- Authentication with smartphones (unlock screen, access to sensitive information etc.)
  - Almost **50% of phones** are **spoofed** by a **photo** (2019)
  - PVID regulation (2021) : video as identity proof.
- Digital sentry in China to fight pandemic (2022) :
  - Face mask detection
  - Temperature control
  - Verification of the vaccination record
  - Identity verification

An illustration of « digital sentry ». (Credit: SenseTime)

- ***Facial Recognition Goes to War (2022)***

Ukraine is receiving free access to Clearview AI's powerful search engine for faces (BusinessToday.In)

Peter Kulche. (2019). Face recognition on smartphone is not always safe

4

Human visual inspections vs. Automatic Facial Recognition

- **Human** perception is related to the **real world** while the perception of the **machine** is linked to the **digital world;**
- The human and the machine have relative differences in the faculty to recognize faces;

	Recognize a face on a "distorted" image	Recognize a face after 20 years	Match first time seen faces
Human	**Good**	**Good**	**Bad**
Machine	**Bad**	**Bad**	**Good**

5

Identify a new face among others

- Is this person in the array?
- If present, match the person.

6

V. Bruce et al., "Verification of Face Identities From Images Captured on Video", (1999)

Identify a new face among others

- Is this person in the array?
- If present, match the person.

7

V. Bruce et al., "Verification of Face Identities From Images Captured on Video", (1999)

Identify a new face among others

- Is this person in the array?
- If present, match the person.

8

V. Bruce et al., "Verification of Face Identities From Images Captured on Video", (1999)

Identify a new face among others

- Is this person in the array?
- If present, match the person.

Not in the array!

9

V. Bruce et al., "Verification of Face Identities From Images Captured on Video", (1999)

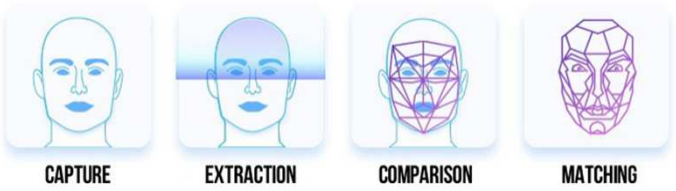
Identify a new face among others: Experiment result

- When target was present in the array. **12%** picked **wrong** person and **18%** said they were **not present** (overall only **70% correct**).
- When target was **not present** in the array **70%** still matched the target to someone in the array.

10

V. Bruce et al., "Verification of Face Identities From Images Captured on Video", (1999)

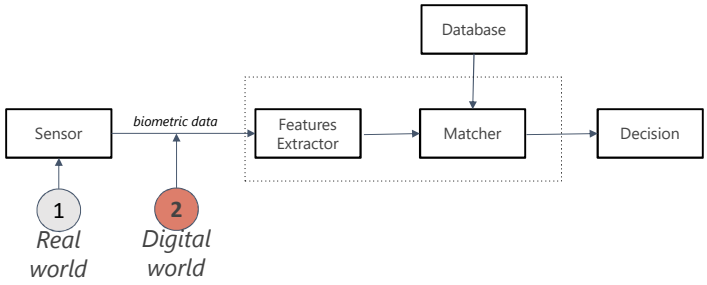
Facial Recognition System



Tecsynt Solutions. (2017, Oct 05). Things You Were Afraid To Ask: Pros and Cons of Facial Recognition Technology For Your Business.

11

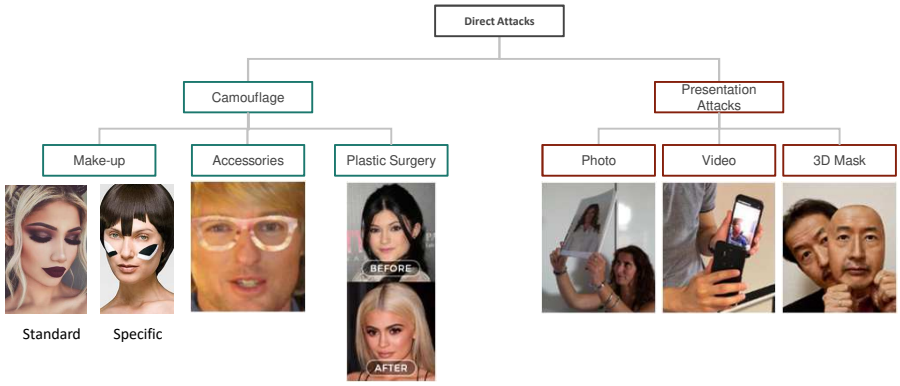
Attacks on facial recognition systems



1. Direct attacks: Presentation attacks (Spoofing attacks and Evasion attacks)
2. Indirect attacks: Inject a fake image after the sensor and before the process

12

Direct attacks



Credits: (*Left to right*) Eurecom, Surys, real-f, @leta_konstantinova, cvdazzle, M.Sharif et al., "A General Framework for Adversarial Examples with Objectives". NextFeed.

13

Facial Cosmetics Database and Impact Analysis on AFR

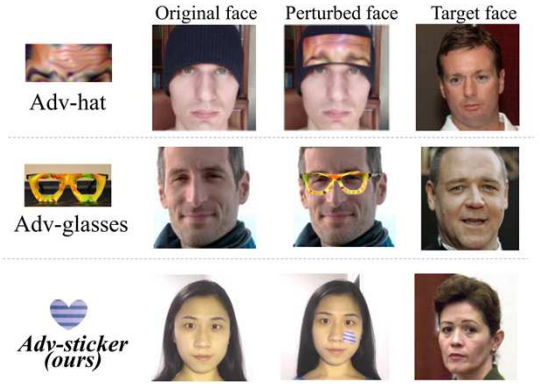
Reference Image and Makeup Series Images in their Original Form

- If the reference images do not include any make-up, facial cosmetics do have a negative impact on automatic face recognition;
- Surprisingly, the application of facial cosmetics in the reference images can help to obtain better identification rates;
- The eye subarea plays a major role in face recognition while the mouth area is not pivotal.

Eckert, Marie-Lena, Neslihan Kose, and Jean-Luc Dugelay. "Facial cosmetics database and impact analysis on automatic face recognition." *2013 IEEE 15th international workshop on multimedia signal processing (MMSP)*. IEEE, 2013.

14

Discredit a system with accessories



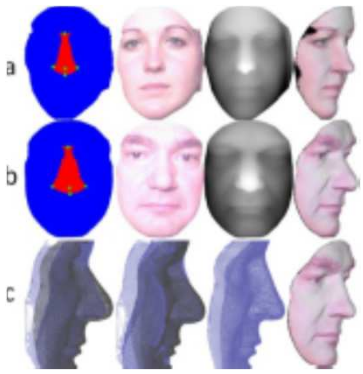
- The idea is to use adversarial attack principle to fool facial recognition software by constructing adversarial objects
- In this way it is possible to be undetectable by facial recognition systems or impersonate someone else

Y.Guo et al., "Meaningful Adversarial Stickers for Face Recognition in Physical World", 2021.
https://doi.org/10.52843/cassyni.bcgygr

15

Plastic surgery

Both 2D & 3D
standard recognition methods
are not robust to the variations caused
by nose alterations, especially for the
verification case.



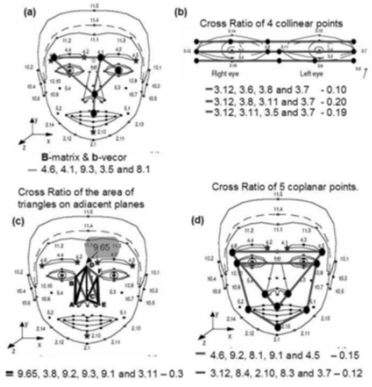
Kose, Neslihan, Nesli Erdogmus, and Jean-Luc Dugelay. "Block based face recognition approach robust to nose alterations."
2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2012.

16

photo

17

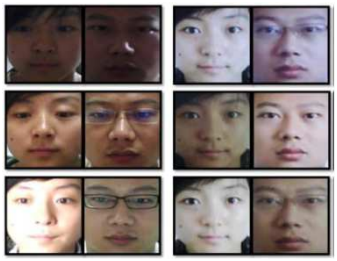
Anti-spoofing



Riccio, Daniel, and Jean-Luc Dugelay. "Geometric invariants for 2D/3D face recognition." *Pattern Recognition Letters* 28.14 (2007): 1907-1914.

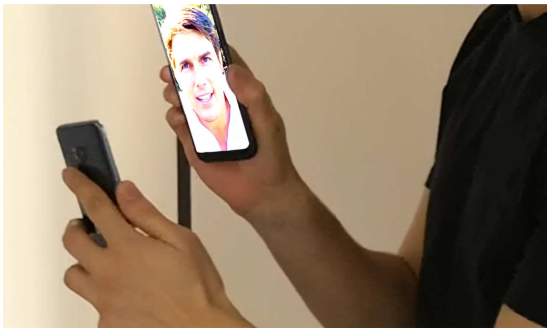
Each column contains samples from session 1, 2 and3.

In each row, the left pair is from a live human and the right from a photo.



Kose, Neslihan, and Jean-Luc Dugelay. "Classification of captured and recaptured images to detect photograph spoofing." *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 2012.

video



- Images recaptured from an LCD screen can leave traces

(aliasing, blurriness, noise);

- Traces are much less present on recaptured images from an OLED display.

Trabelsi, Anis, Marc Pic, and Jean-Luc Dugelay. "Recapture Detection to Fight Deep Identity Theft."
Proceedings of the 2022 4th International Conference on Video, Signal and Image Processing. 2022.

19

3D

20

3D / Spoofing ethnicity

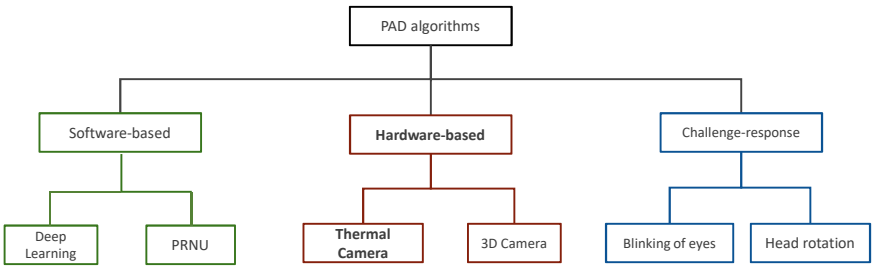
In 1999, Henley Stephenson (left) stole a bank in UK with a mask of a white face

In 2010, Conrad Zdziera (right) stole a bank in USA with a Super Realistic Hollywood Mask of a black face

21

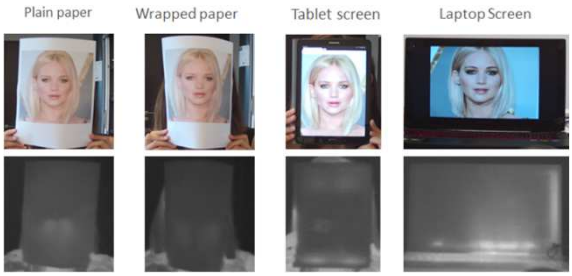
Presentation Attack Detection (PAD) Algorithms

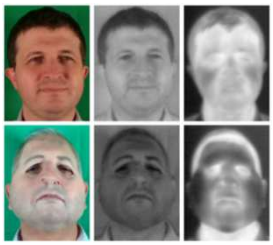
- Software-based : First way is to use available sensors and to detect in the signal a pattern characteristic of liveness/spoofing.
- Hardware-based : Second method is to use dedicated sensors to detect an evidence of liveness, which is not always possible to deploy.
- Challenge-response : Liveness detection by asking the user to interact with the system.



22

Thermal Imagery is commonly known as a natural spoofing countermeasure

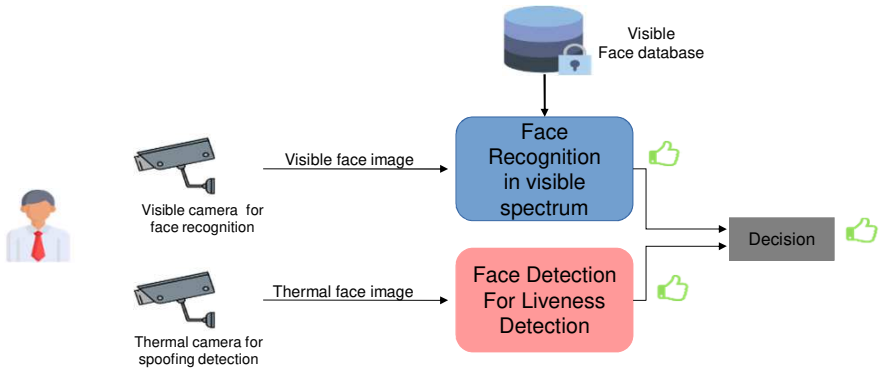




Lei Li et al., "Face recognition under spoofing attacks: countermeasures and research directions", 2018.
S. Marcel et al., "Spoofing Deep Face Recognition with Custom Silicone Masks", 2018.

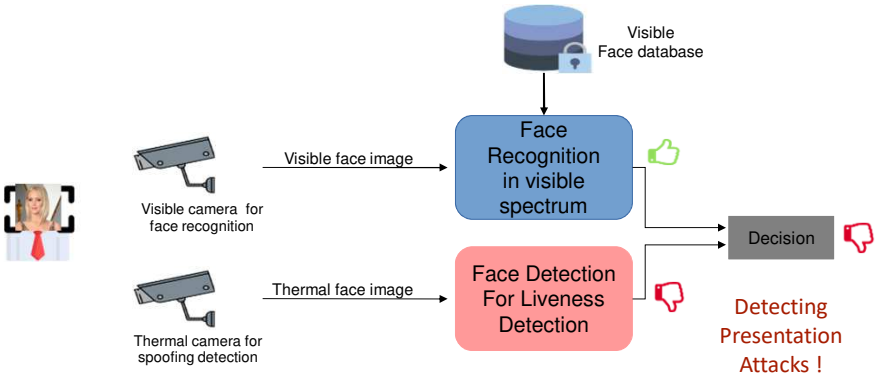
23

Thermal Camera mechanism



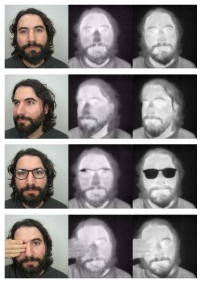
24

Thermal Camera mechanism



25

Create synthesis visible image from thermal image

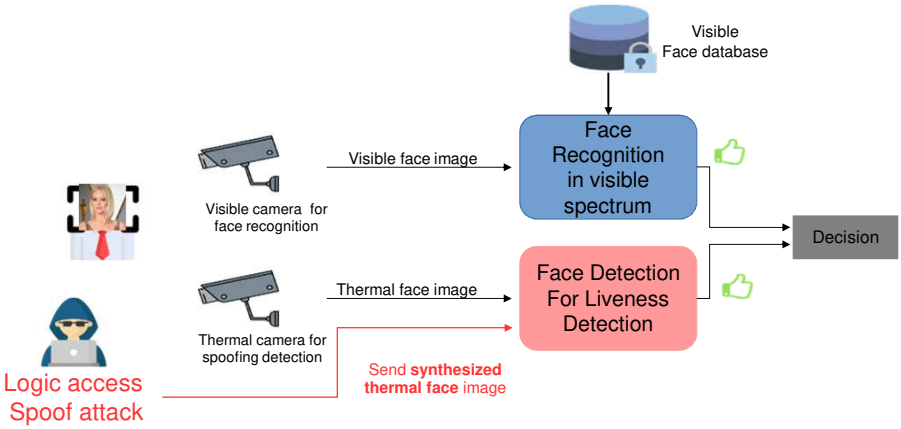


- Generating thermal-like images from visible captures
- Synthetic images can be used to spoof the system

Mallat, Khawla, and Jean-Luc Dugelay. "Indirect synthetic attack on thermal face biometric systems via visible-to-thermal spectrum conversion."

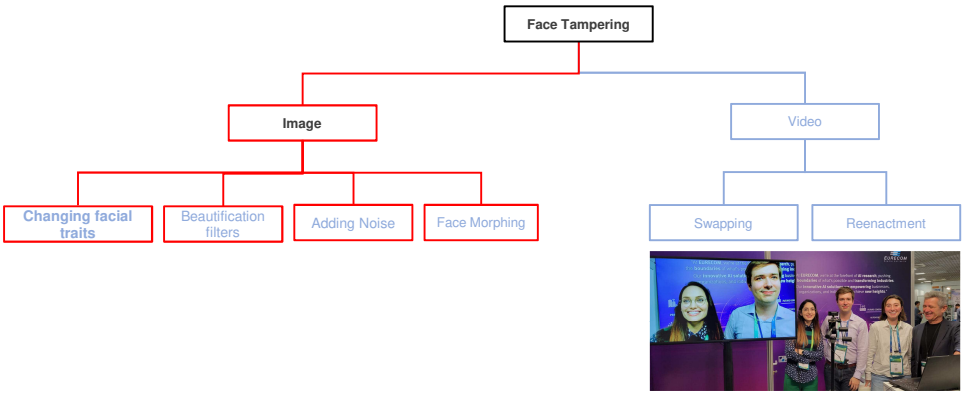
26

Thermal Camera mechanism / Attack through the digital world!



27

Face tampering techniques (in the digital world)



28

Changing Facial Traits; *e.g. FaceApp*

Popular application (FaceApp): Allows to changing facial traits (eyes color, hair color, gender etc.) in a realistic ways (can fool human visual inspection!).

Removing beard

29

Z. He et al., "AttGAN: Facial Attribute Editing by Only Changing What You Want", (2017)

Changing facial Traits; e.g. Age tampering: Rejuvenation & Aging

Original image

- With Deep Learning it's possible to create new data from existing ones
- Create with accuracy the face photo of a person at 50 years old knowing his/her picture at 30

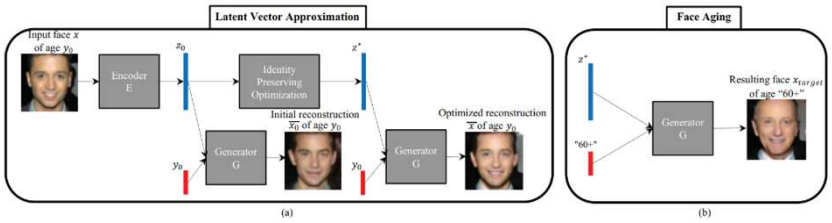
Younger Older

Synthetic images

30

Face aging

Framework based on conditional GAN (Generative Adversarial network) in order to modify the apparent age of an individual.



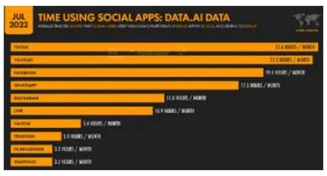
(a) approximation of the latent vector to reconstruct the input image
(b) switching the age condition at the input of the generator G to perform face aging

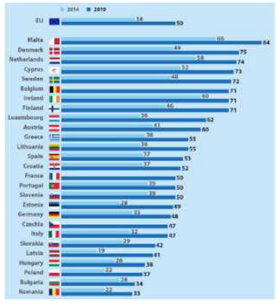
Antipov, Grigory, Moez Baccouche, and Jean-Luc Dugelay. "Face aging with conditional generative adversarial networks." *IEEE international conference on image processing (ICIP)*. IEEE, 2017.

31

Beautification filters: introduction

Beautification is the process of making visual alterations to the perceived shape and texture of a human face in order to increase the beauty of the subject.



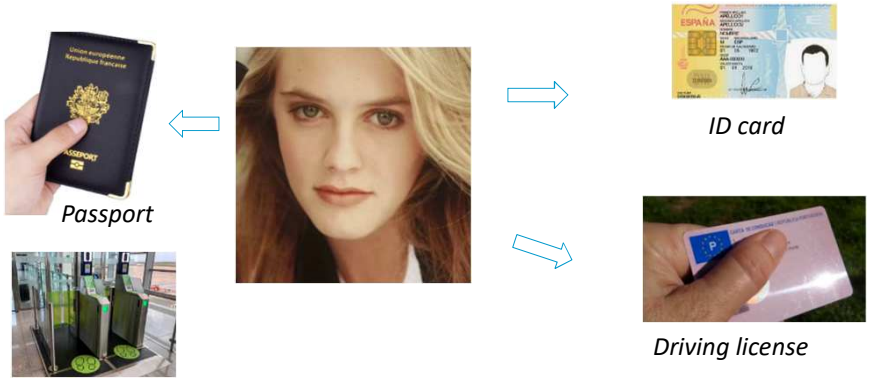


- Social media platforms offer different filters to beautify images.
- Filtered images are some of the most heavily engaged photos on social media

LAVRENCE, Christine et CAMBRE, Carolina. "Do I look like my selfie?": Filters and the digital-forensic gaze. Social Media+ Society, 2020, vol. 6, no 4, p. 2056305120955182.

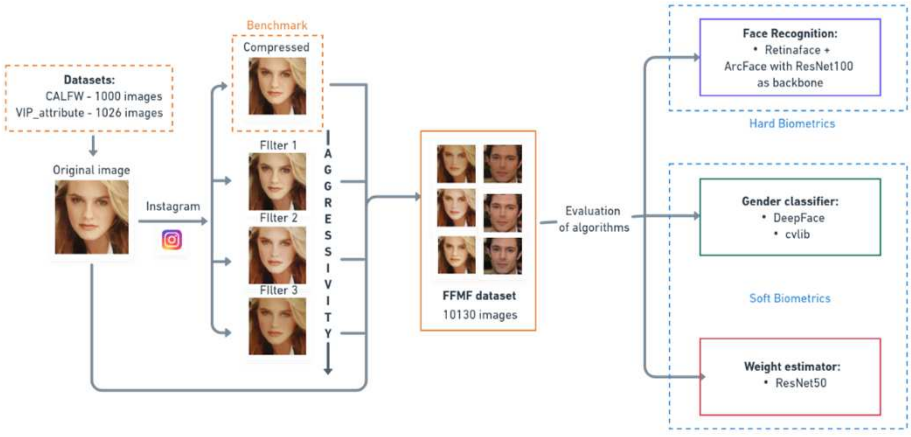
32

Beautification filters: usages



33

Beautification filters: impact on FR



Mirabet-Herranz, Nelida, Chiara Galdi, and Jean-Luc Dugelay. "Impact of Digital Face Beautification in Biometrics." *EUVIP 2022, 10th European Workshop on Visual Information Processing*. 2022.

34

Beautification filters: experimental results

- Face Verification
- Gender classification
- Weight estimation

AGGRESSIVITY

HIGHER SECURITY

Verification	Original	Compressed	Thinner_face	Relax! You Pretty!	Glam grain
10	12.38	12.98	12.47	11.98	14.05
1	17.63	17.35	17.30	15.50	21.38
0.1	21.98	21.11	25.53	22.42	32.23

Gender	All	Female	Male
Original	91,81	84,40	**99,22**
Compressed	86,84	75,63	98,05
Thinner_face	87,13	76,02	98,24
RelaxYouPretty	87,62	79,14	96,10
Glam grain	**93,56**	**88,10**	99,02

Accuracy: the higher the better

Weight	Original	Compressed	Thinner_face	Relax! You Pretty	Glam grain
MAE	**8,52**	8,65	8,69	8,89	9,16
Me	2,79	2,55	**2,42**	3,92	2,86
STDe	22,82	22,94	23	22,78	**22,42**

Error metrics: the lower the better

35

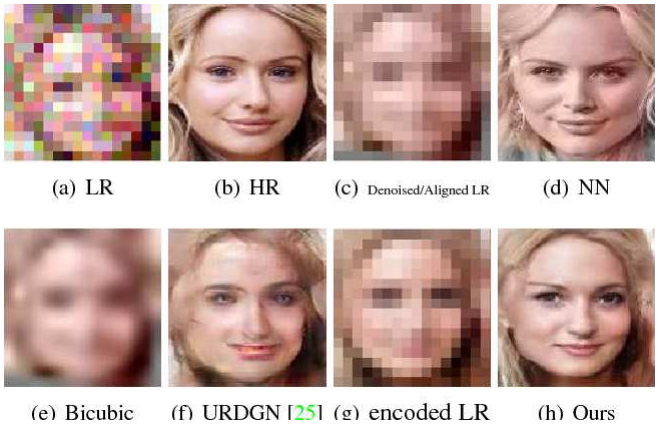
Face (de)Anonymization

In video surveillance it's common to protect privacy of people (anonymize face).

Three methods of anonymization
(left to right: blurring, pixelization, black masking)

36

Hallucinating

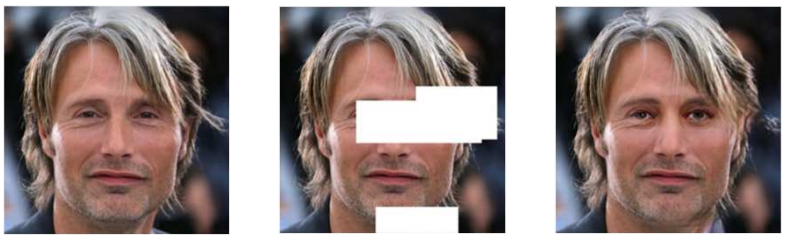


37

X Yu et al., "Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders", 2018.

Face reconstruction

It is possible to reconstruct a face after partial "covering" the face



Left: Original image, Middle: Mask, Right: Reconstruction

38

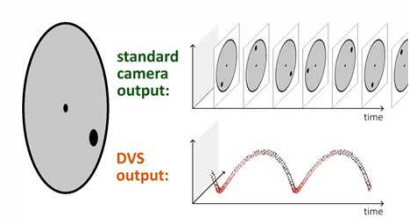
G.Liu et al., "Image Inpainting for Irregular Holes Using Partial Convolutions", 2018.

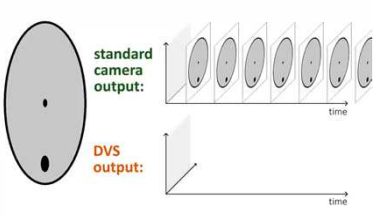
Event-based camera

Event camera generates asynchronous sequence of events

Advantages:

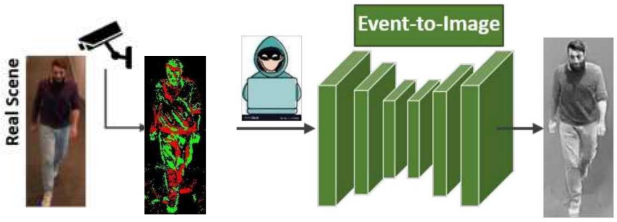
- High dynamic range (140 dB instead of 60 dB)
- Low latency (~ 1 μs)
- Energy efficient





39

Inversion Attack



40

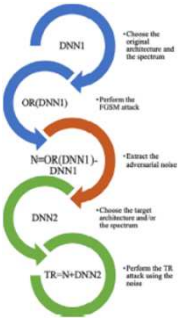
S. Ahmad et al., "Person Re-Identification without Identification via Event Anonymization", (2023).

Adversarial attacks on machine learning models

Example for face recognition

41

Adversarial attacks through architectures and spectra in face recognition



Bisogni, Carmen, et al. "Adversarial attacks through architectures and spectra in face recognition." *Pattern Recognition Letters* 147 (2021): 55-62.

42

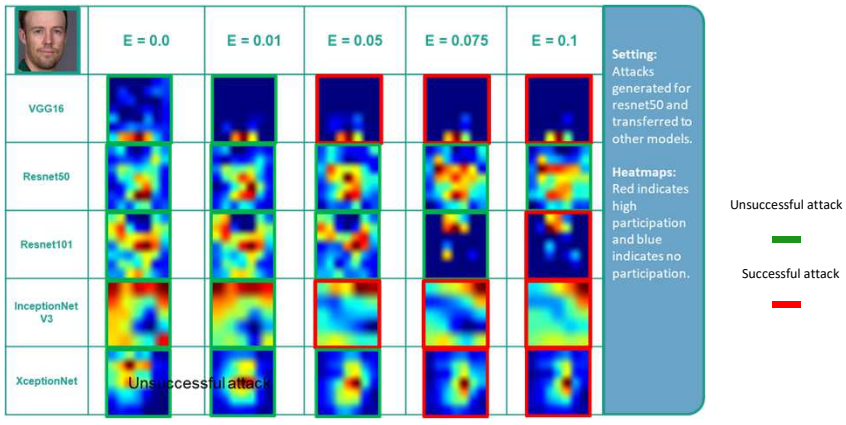
Face recognition model under Adversarial attack

Models	Ɛ = 0.000	Ɛ = 0.010	Ɛ = 0.050	Ɛ = 0.075	Ɛ = 0.100
VGG 16	87.54%	85.60%	85.21%	84.82%	84.43%
Resnet50	94.16%	94.55%	90.66%	85.21%	75.09%
Resnet101	93.38%	93.38%	87.15%	84.82%	77.43%
InceptionNet v3	85.21%	84.82%	66.92%	47.85%	29.96%
XceptionNet	89.49%	90.27%	71.20%	59.14%	45.13%

Accuracy of different models under FGSM adversarial attack

43

Shifting Behavior in the region of Participation in CNN



44

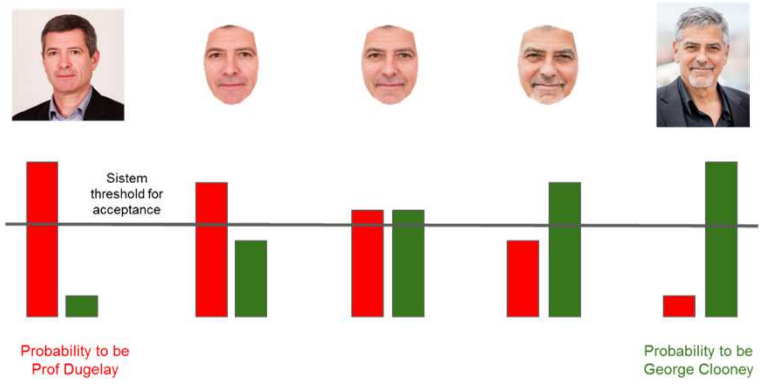
CNN under Adversarial Attack

Shifting behavior in the region of participation in CNN		
Adversarial perturbation not perceptible to the human eye but changes model decision	Deviation of participation in decision making region	Non-deterministic shifting

Chakraborty, Tanmay, et al. "Generalizing adversarial explanations with Grad-CAM."
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

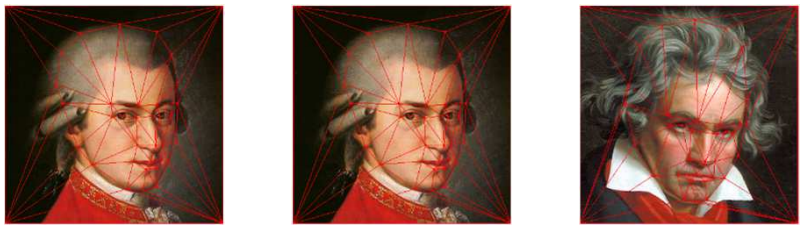
45

Face Morphing Attack: Principle



46

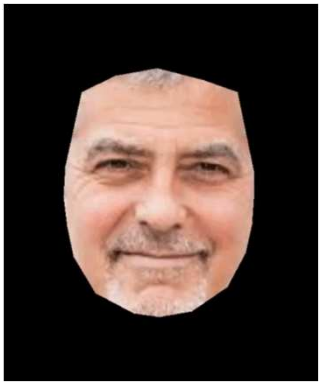
Face Morphing - Steven Traversi. (2016).



1. Identify feature points on both images
2. Create a triangulation from one of the sets of points, and apply it to each image.
3. Apply a affine transform

47

Morphing Attack: Video

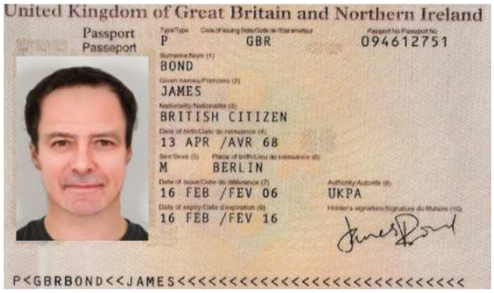


48

The code used to obtain the displayed image is written in Python and can be downloaded from: https://github.com/alyssaq/face_morpher

Face Morphing; Application: The Magic Passeport

- What if was it possible that a face image can belong to two different subjects at the same time?
- That would allow two persons to use the same passport (for example) without apply any forgery on it.



49

M. Ferrara et al., "The magic passport". (2014)

Ultimately…

50

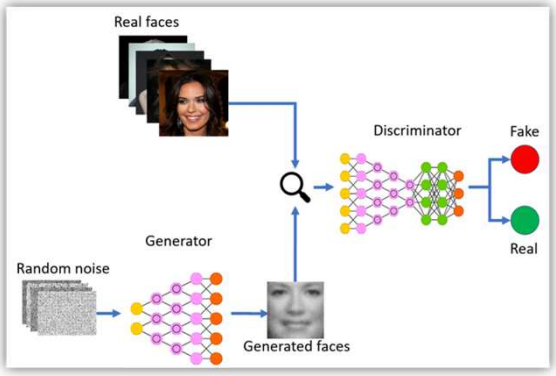
whichfaceisreal.com (2019)

Ultimately…

whichfaceisreal.com (2019)

51

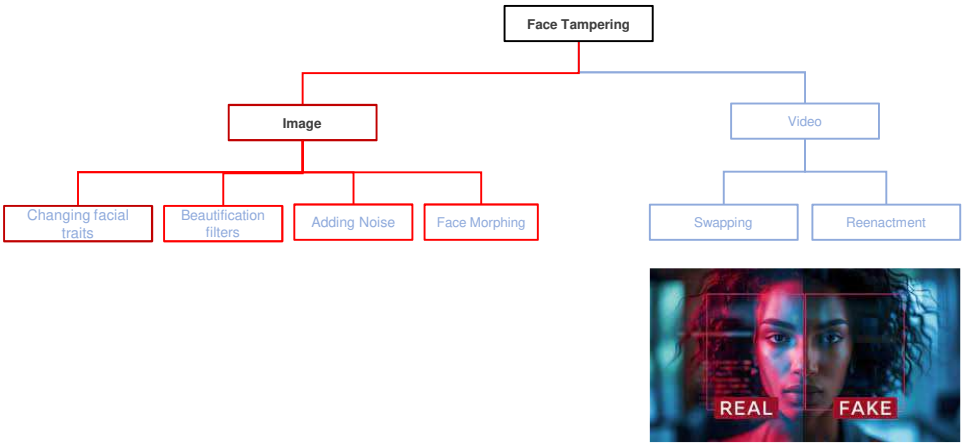
GAN … Artificial faces… *In just a short few years…*



Introduced by Ian Goodfellows in 2014
Aim to synthesize artificial samples (like images)
The GAN images have become so realistic that they can fool a human viewer.
Ian Goodfellow (2019) « 4.5 years of GAN progress on face generation ».

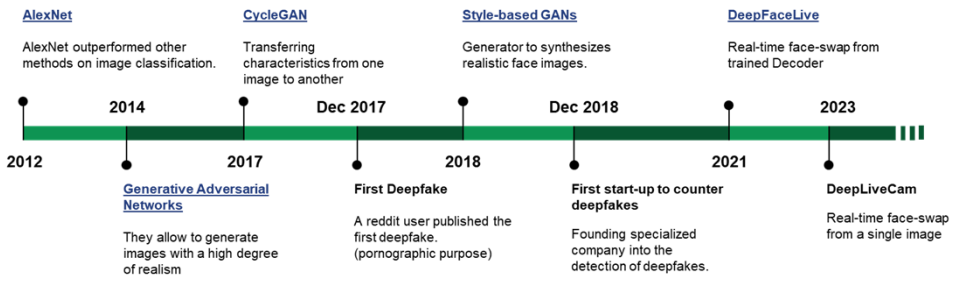
52

Face tampering techniques (in the digital world)



53

Short Deep Learning Timeline



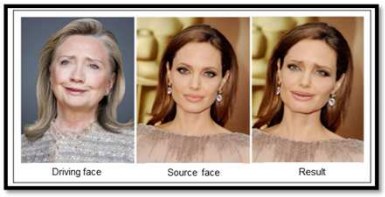
- Deep Learning brought uncertainty in our beliving of visual evidence;
- Deep Learning has a central role in both Fake Generation and Fake Detection.

54

Face-swap & Face-reenactment

The face of the source image is replaced with the face of the target image. The content transferred from source image to target image is driven by target image

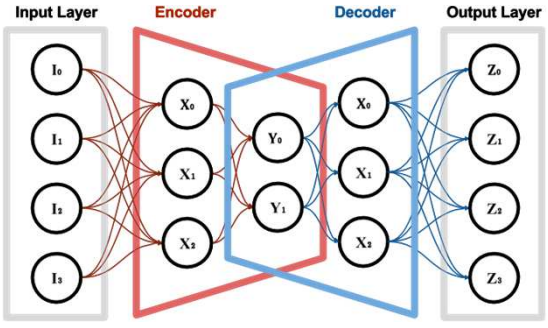
Targeting image/video frame is used to drive the expression, gaze , pose, or body of the source image. Turns an identity into a puppet

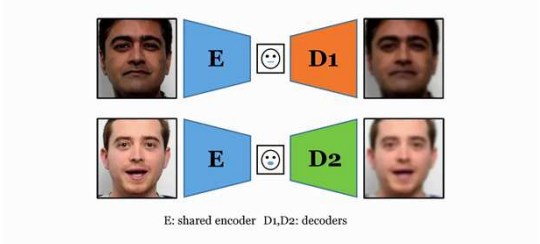


Masood, Momina, et al. "Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward." Applied Intelligence 53.4 (2023): 3974-4026.

55

Auto-encoders





E: shared encoder D1,D2: decoders

56

First Order Motion Model

Swap

One single image (target) – Source video – Resulting fake video
* Unseen people (swap) *

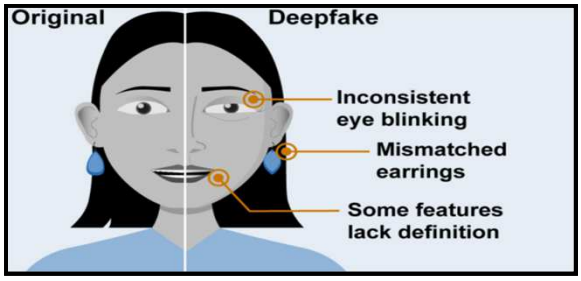
Reenactment

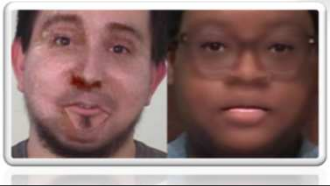
One single image (target) – Source video – Resulting fake video
* Unseen people (reenactment) *

57

A. Siarohin et al., "First Order Motion Model for Image Animation", (2019).

How can you spot a deepfake?





58

DeepFake Detection Challenge (DFDC)

Final rank	Results	Our re-implementation
1st	0.1983	0.1957
2nd	0.1787	0.1790
3rd	0.1703	0.1821
4th	0.1882	0.1863
5th	0.2157	0.2158

Eq.1 $LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$

Dataset	# Fake videos	# Real videos	# identity	# methods	# augmentation
UADFV (2018)	49	49	49	1	-
DeepfakeTIMIT (2018)	640	320	43	2	-
FaceForensics++ (2019)	4000	1000	?	4	2
Google DFD (2019)	3000	363	28	5	-
Celeb-DFD (2019)	5639	890	59	1	-
DeeperForensics-1.0 (2020)	1000	59000	100	1	7
DFDC (2020)	104500	23654	960	5	19

	1st	2nd	3rd	4th	5th
1st	100 %	48 %	47%	49%	27%
2nd		100 %	48%	41%	26%
3rd			100%	53%	22%
4th				100 %	27%
5th					100 %

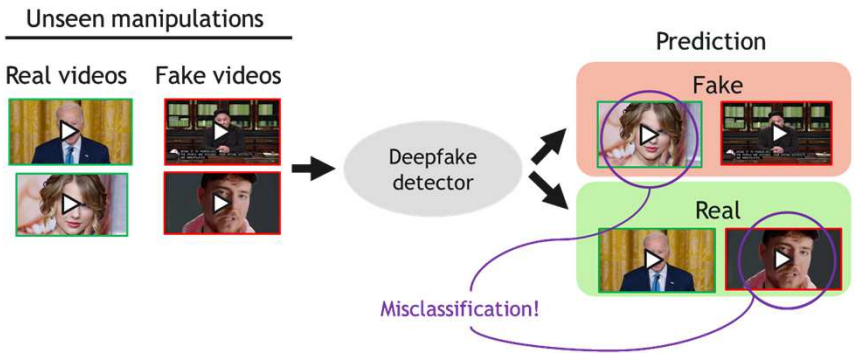
The best solution that was proposed during this challenge obtains the accuracy of 82%, but drops to 65%
when testing randomly taken videos from the internet

A simple deep ensemble among top-5 solutions enables to improve the results by 20%

Trabelsi, Anis, Marc Michel Pic, and Jean-Luc Dugelay. "Improving Deepfake Detection by Mixing Top Solutions of the DFDC." *2022 30th European Signal Processing Conference (EUSIPCO).*

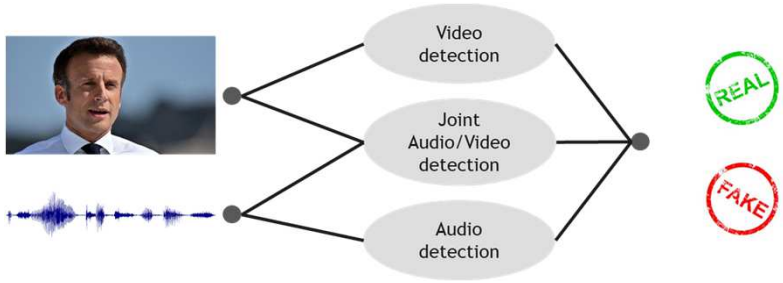
59

DeepFake detectors are not robust to unseen manipulation methods



60

DeTOX: Fight against deepfakes of French personalities



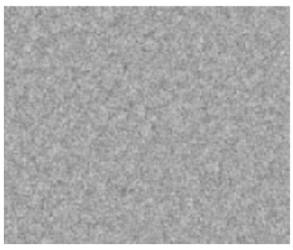
61

French National Project DeTOX: http://detox.eurecom.fr

Image Forensics - PRNU

PRNU is a distinctive pattern due to imperfections in the silicon wafer, during the sensor manufacturing, even among cameras of the same model/same brand.

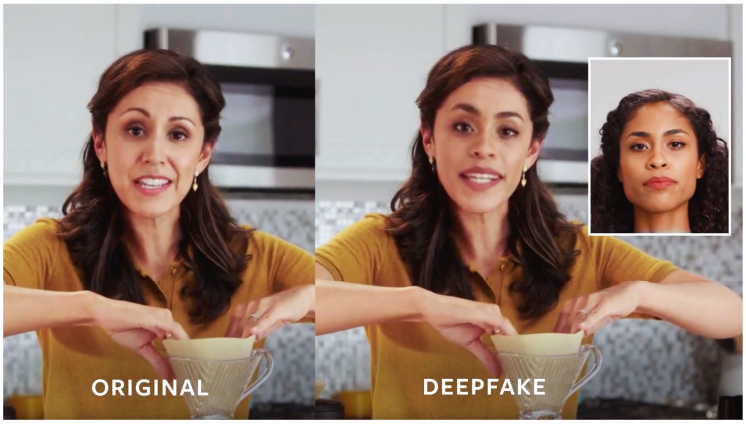
It is extremely unlikely that 2 sensors have the same PRNU pattern.

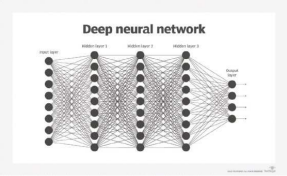


62

Digital watermarking?

allows **to hide** an **invisible** and **robust** message inside audio, image, and video.



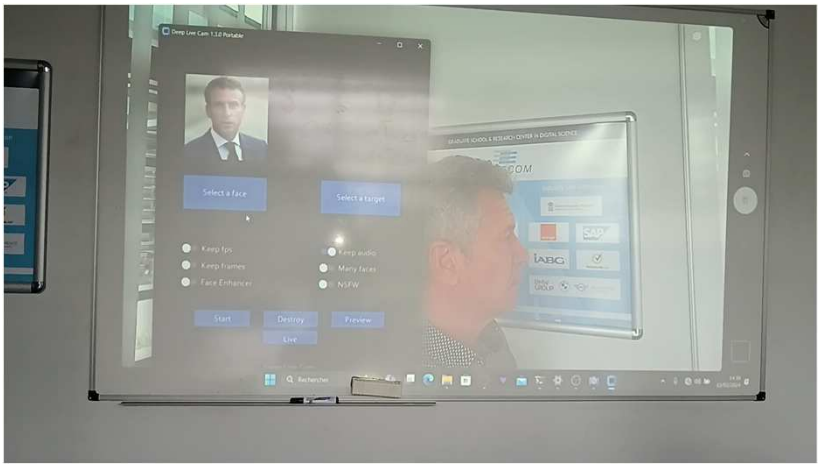


63

DeepLiveCam (reference)

64

DeepLiveCam (target)



65

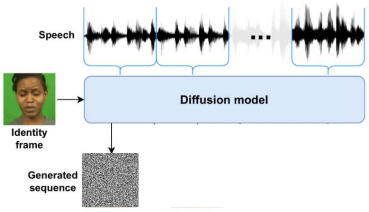
DeepLiveCam (deepfake in real time without any training)

66

To conclude…. and what's about Speech…

67

Audio-driven diffusion-based deepfakes



Output Example

68

Stypułkowski et al., "Diffused Heads: Diffusion Models Beat GANs on Talking Face Generation", (2023).

Thank you for your attention
Seeing is (no longer) believing

69