Media Forensics and the challenge of Deepfakes
in a social media context

Irene Amerini
amerini@diag.uniroma1.it

Luca Maiano
maiano@diag.uniroma1.it

19/05/2021

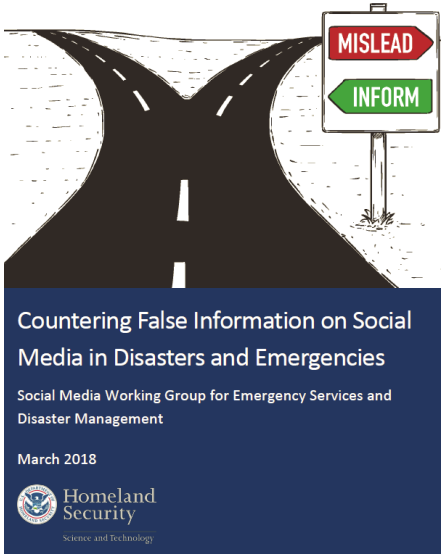
SAPIENZA
UNIVERSITÀ DI ROMA

Can you trust what you see?

- Multimedia contents have become increasingly important in everyday life. The expressiveness of visual content makes multimedia a powerful means of communication, however, it becomes increasingly important to be able to verify the source and authenticity of this information.

The contexts

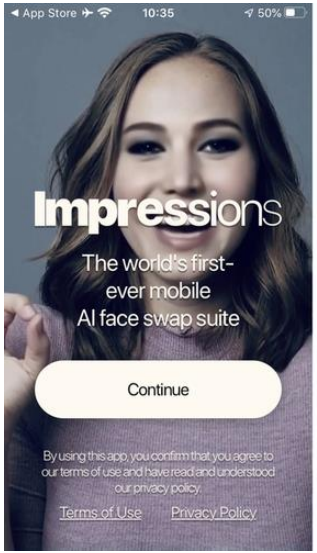
- Weaponized information and information warfare, where the organic propagation of virulent misinformation is under analysis.
- Propaganda and military purposes
- In a court of law (reputation attacks, document frauds, crime scene alterations)



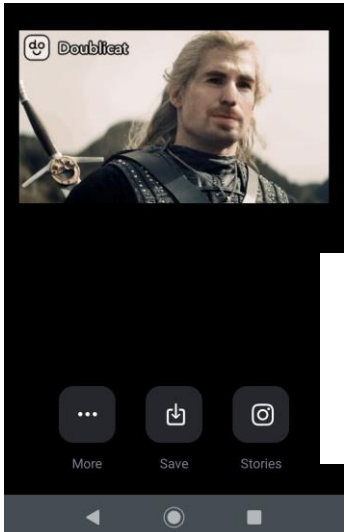
A new threat

- Deep Fakes phenomena with AI
- Deepfake videos are AI-generated realistic sequences

reddit



ZAO

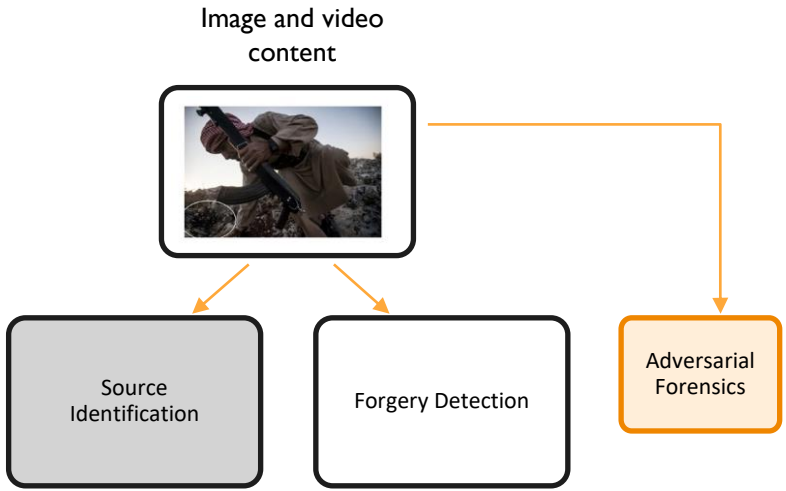


https://beebom.com/best-deepfake-apps-websites/

Image and Video Forensics

Multimedia forensics aims to answer such challenges with techniques that are capable to assess:

- **Authenticity**
  - Forgery detection, i.e. deciding on the integrity of the media.
  - Deepfakes detection; i.e. artificially generated content.
- **Origin**
  - Source identification, i.e. link multimedia content to a particular device or social network.
- **Security**
  - Adversarial forensics/Counter forensics

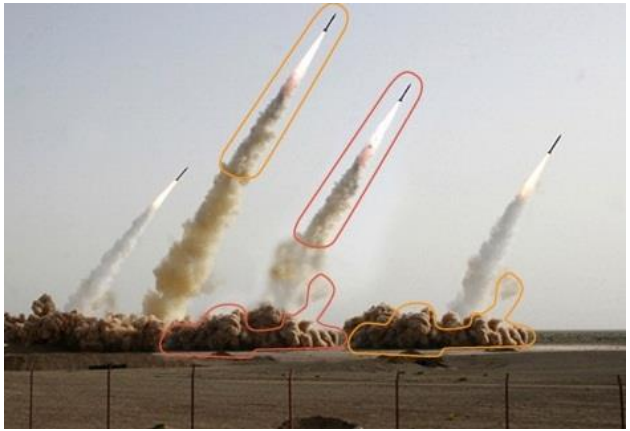


Kinds of manipulations

- Image manipulation categories:
  - Image Splicing
  - Copy-Move manipulation
  - Deepfakes

Kinds of manipulations

- Image manipulation categories:
  - Image splicing
  - Copy-Move manipulation
  - Deepfakes

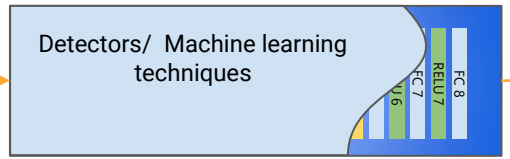


Kinds of manipulations

- Image manipulation categories:
  - Image splicing
  - Copy-Move manipulation
  - Deepfakes

Forgery detection

- **Research question:** how a doctored image/video be revealed and localized?
- Given a single probe image, detect if the probe was manipulated or artificially generated and provide mask(s)



The image is doctored with a certain confidence

Source identification

Which **CLASS** of devices

Which **BRAND/MODEL**

Nikon?
Canon?
Sony?

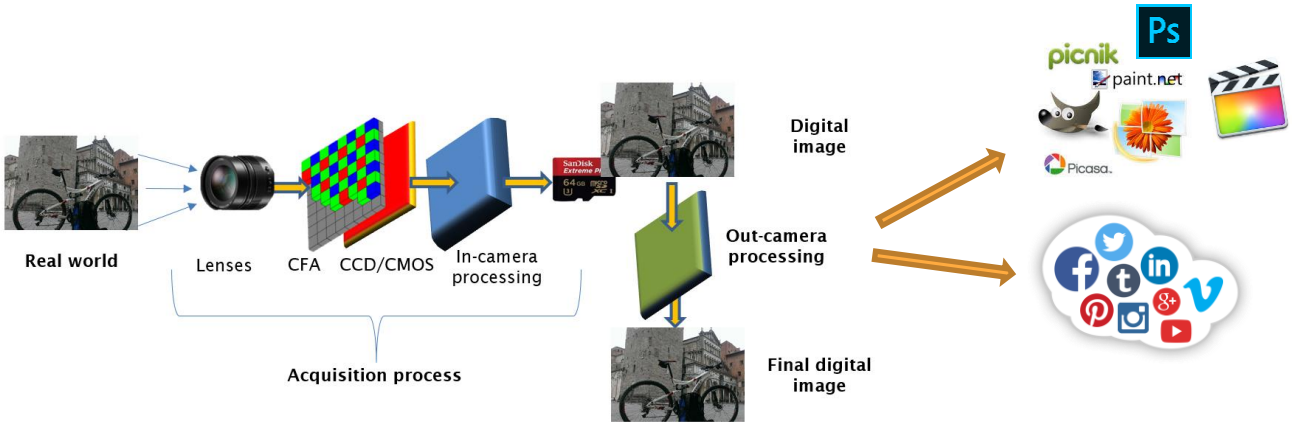
Which **DEVICE**

Which Nikon D3300?

Which sharing platform?

Basic principles 1

- Each processing/manipulation leaves on the media peculiar traces that can be exploited to make an assessment on the content itself.
- Image and video forensic techniques gather information on the history of images and videos contents.



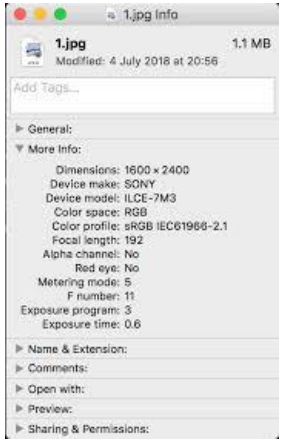
Each phase leaves distinctive footprints!

- at the signal level
- at the metadata/file container level

Basic principles 2

- Acquisition process and post-processing operations leave a distinctive imprint on the data like a **digital fingerprint**
  - *Fingerprint extraction*
  - *Fingerprint classification*

- MM forensics techniques are mostly:
  - **Blind:** original reference media is not required: no side information like metadata
  - **Passive**: different from “active methods” which hide a mark in a picture when it is created like *digital watermarking:* no specific on-device hardware required



Overview on deepfakes

- What are deepfakes?
- Generative models
- Deepfake detection

Let’s play a game –voting link to mentimeter

Go to www.menti.com and use the code 7835 0511

Let’s play a game… which faces are fake?

A

B

C

D

Let’s play a game… which faces are fake?

A

B

C

D

B and C

This Person Does Not Exist

Second chance… which faces are fake?

A

B

C

D

Second chance… which faces are fake?

A

B

C

D

All of them!

This Person Does Not Exist

This person exists… but these are fake!

@deeptomcruise

What are deepfakes?

What are deepfakes?

- Face manipulations in images/videos
- Photoshopped images
- All manipulated images/videos
- Everything in computer graphics
- Any synthetic media: images, videos, text, audio…
- Anything where AI is used (“Deep” from Deep Learning)
- …

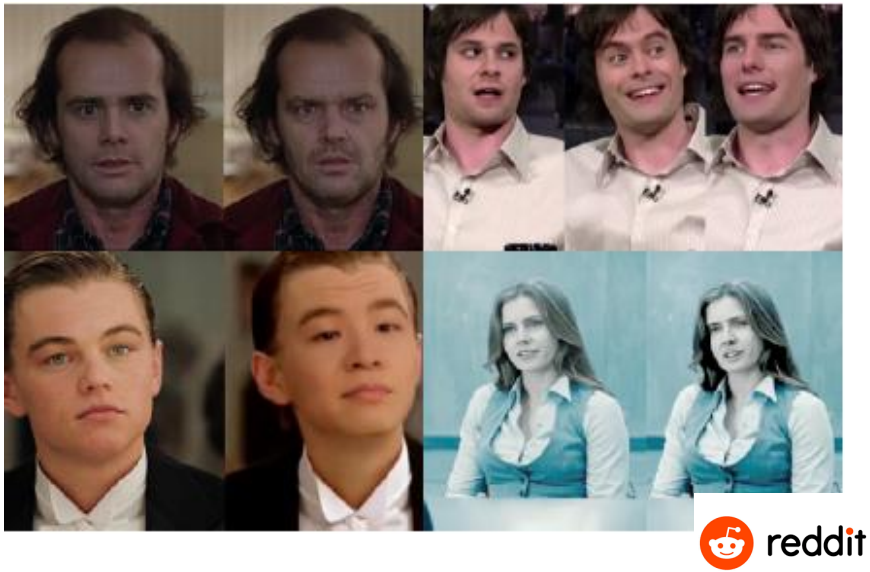
Wikipedia: “Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness.”

A general definition here

Manipulated Images and Videos of Faces

The ‘original’ deepfake

Deepfake are AI-generated realistic images and video sequences



The ‘original’ deepfake method: https://github.com/deepfakes/faceswap

Face Swap vs Reenactment

Identity swap

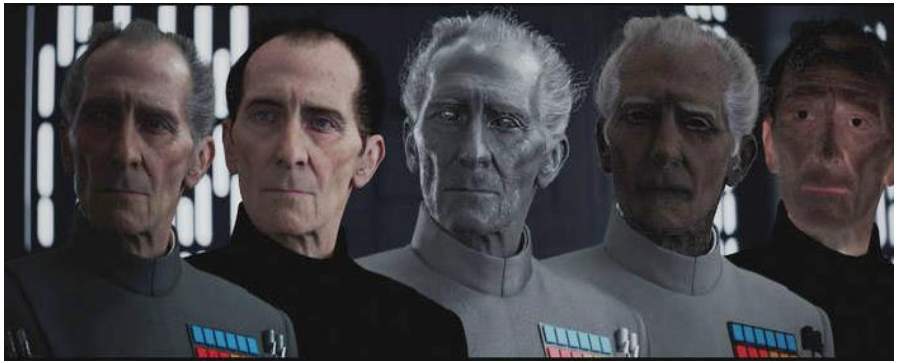
Encoders/
Decoders

Expression
reenactment

Face2Face,
NeuralTextures

Tolsana et al., ’’DeepFakes and Beyond: A Survey of Face Manipulation and Face Detection’’, 2020

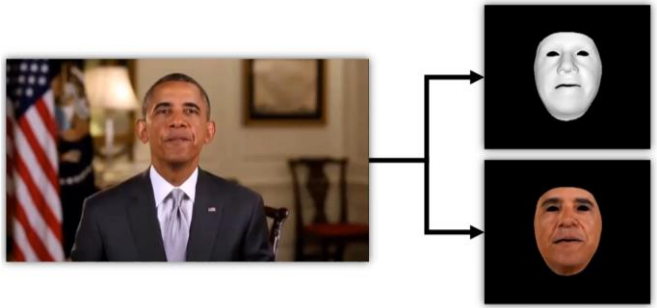
Computer Graphics vs Deep Learning



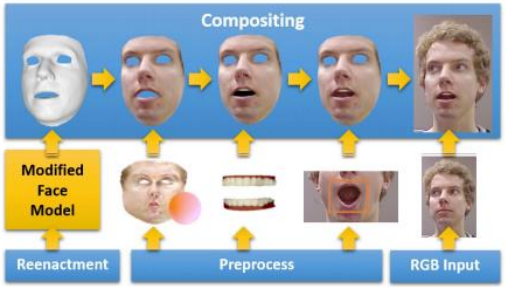
3D Model + Textures + Shading → Syntethic image

Generative models

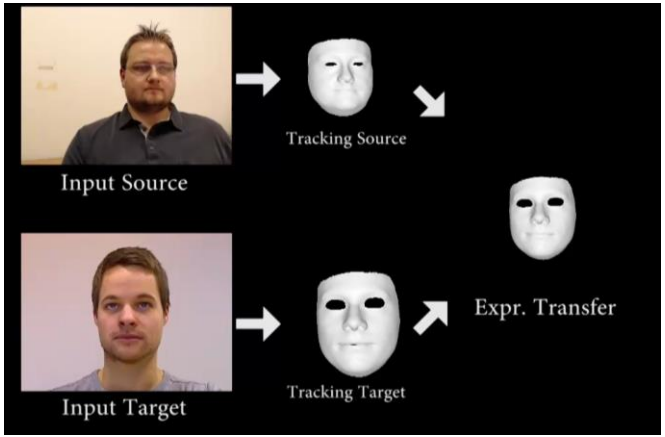
Face2Face



1. Fitting parametric model to RGB image



2. 3D model + image-based rendering



3. Facial expression transfer

Thies et al., “Real-time expression transfer for facial reenactment”, ACM Trans. Graph 2015
Thies et al., “Face2Face: Real-time Face Capture and Reenactment of RGB Videos”, CVPR’16

Facial video editing

- Face Swap vs Reenactment / Video graphics vs Deep Learning (GAN)
- Many techniques: FaceTransfer, Face2Face, DeepFake, Deep Video Portaits, FaceSwap, Neural Textures etc…

[FakeApp, Reddit]

reddit

[Face2Face Niesser et al, CVPR2016]

[FaceSwap]

Applications and risks

- Entertainment, advertising, e-commerce and movie production
- Medicine
- Climate

- Weaponized information
- Manipulation of the public image of celebrities and politicians
- Revenge porn
- Money frauds

Responses

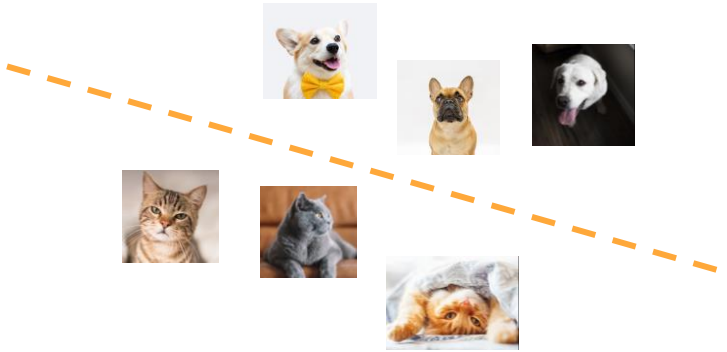
- Social networks bans deepfakes [1]
- YouTube banned deepfakes related to the U.S. presidential election of 2020 [2]
- New laws make it illegal to distribute some deepfakes (like political or porn deepfakes) [3-6]

[1] https://www.theverge.com/2020/1/7/21054504/facebook-instagram-deepfake-ban-videos-nancy-pelosi-congress
[2] https://www.marketwatch.com/story/youtube-to-remove-deepfakes-and-birther-videos-ahead-of-2020-election-2020-02-04
[3] https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/
[4] https://www.businessinsider.com/china-making-deepfakes-illegal-requiring-that-ai-videos-be-marked-2019
[5] https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/
[6] https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce

Generative Models

Discriminative Models vs Generative Models

Discriminative models



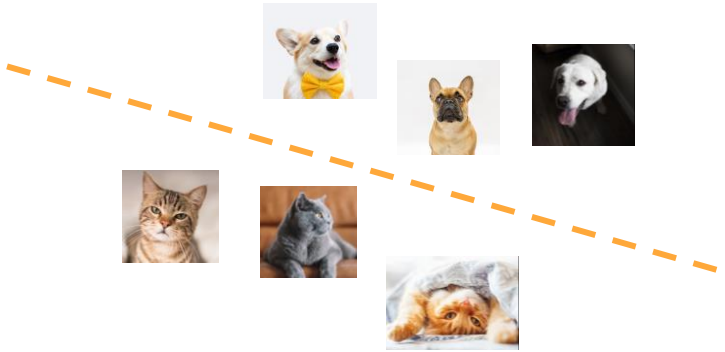
Features Class

$$X \longrightarrow Y$$

$$P(Y|X)$$

Discriminative Models vs Generative Models

Discriminative models



Features Class

$$X \longrightarrow Y$$

$$P(Y|X)$$

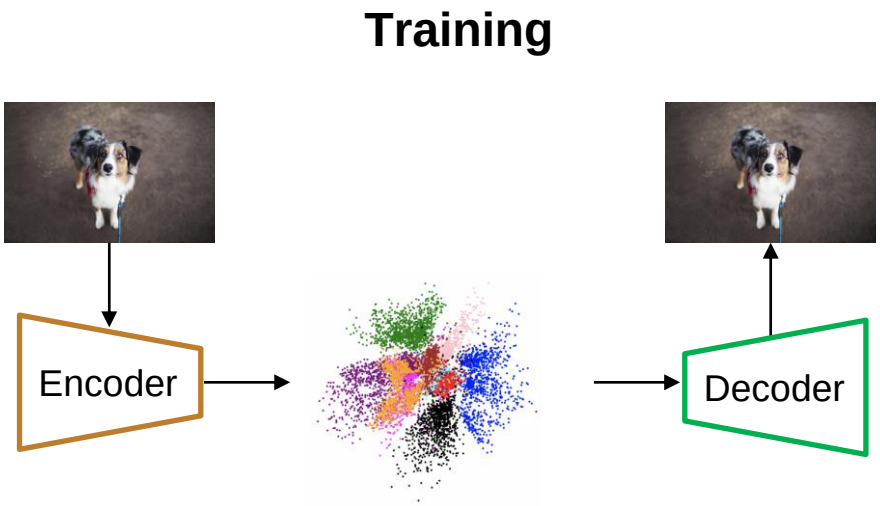
Generative models

Noise, Class Features

$$\xi, Y \longrightarrow X$$

$$P(X|Y)$$

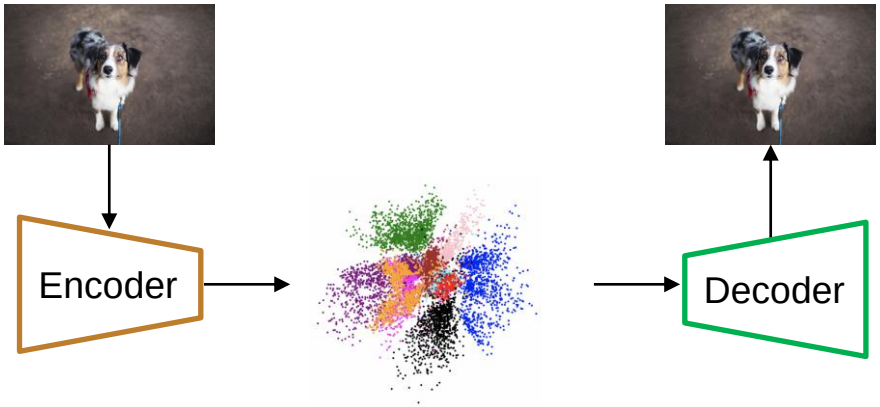
Variational Autoencoders (VAE)



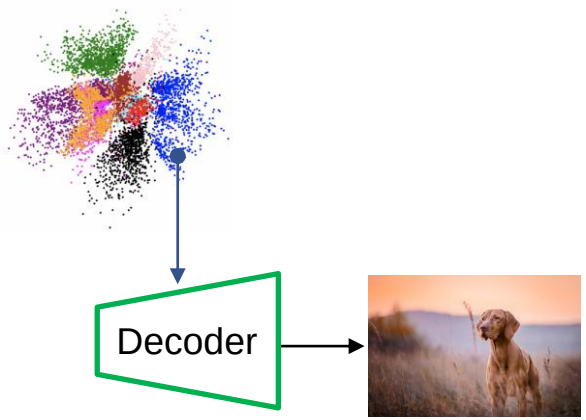
Davinsor et al., “Hyperspherical Variational Auto-Encoders”, UAI-18

Variational Autoencoders (VAE)

Training



Testing



Davinsor et al., “Hyperspherical Variational Auto-Encoders”, UAI-18

Generative Adversarial Networks (GAN)

Training

Real

Discriminator

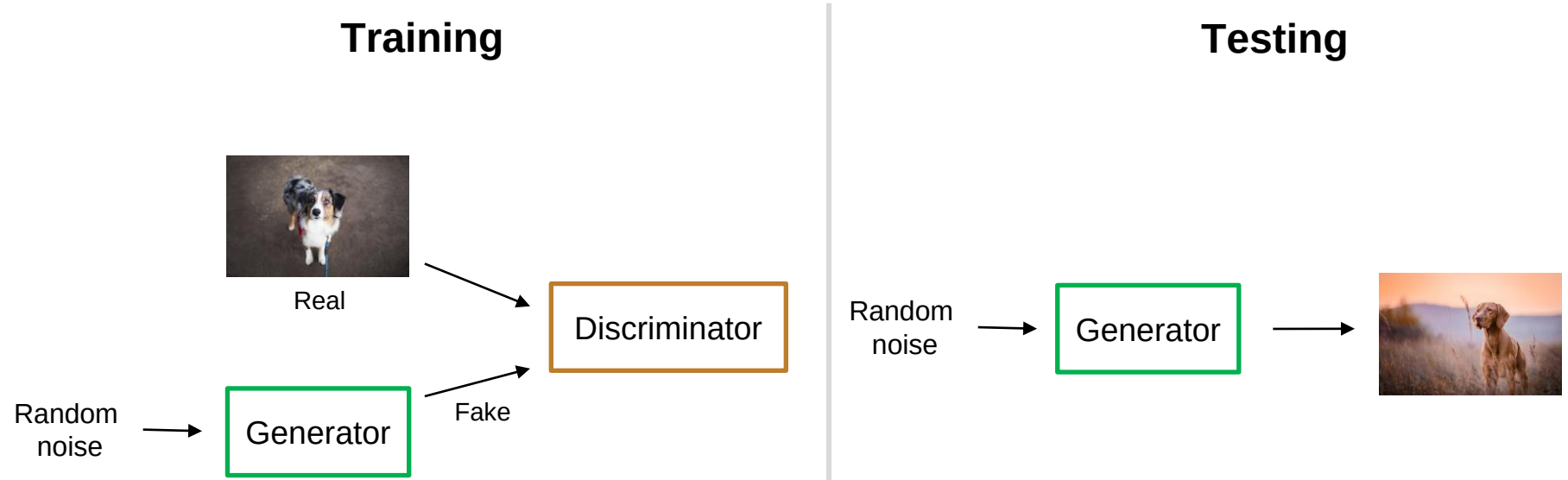
Random noise

Generator

Fake

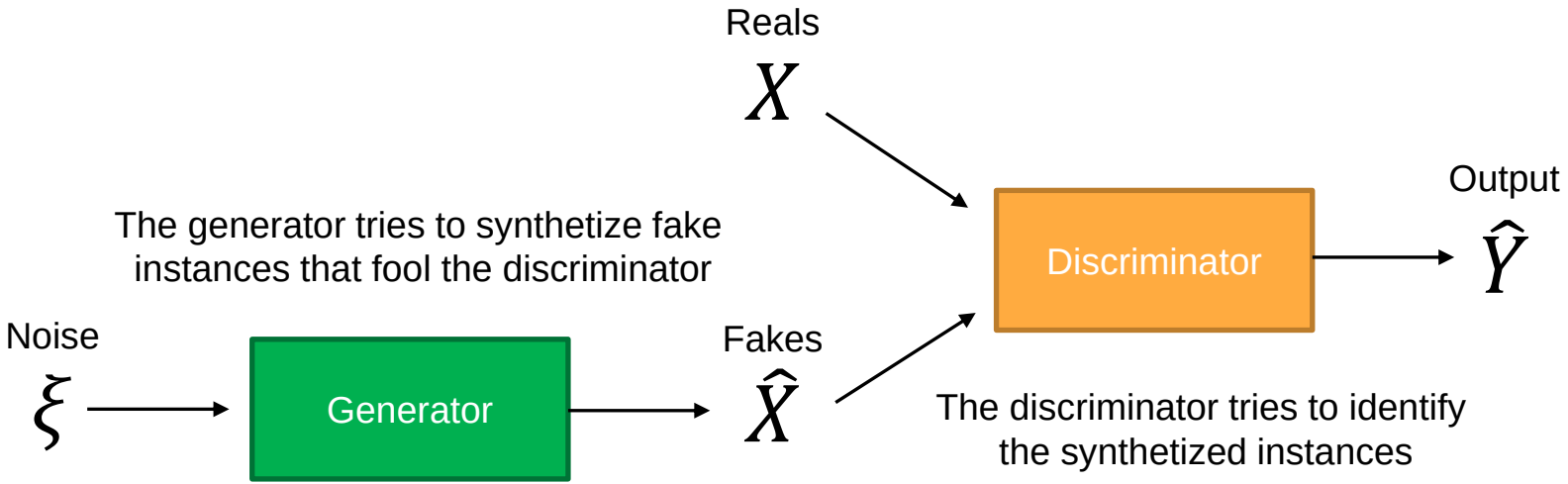
Goodfellow et al., “Generative adversarial nets”, NIPS’14

Generative Adversarial Networks (GAN)



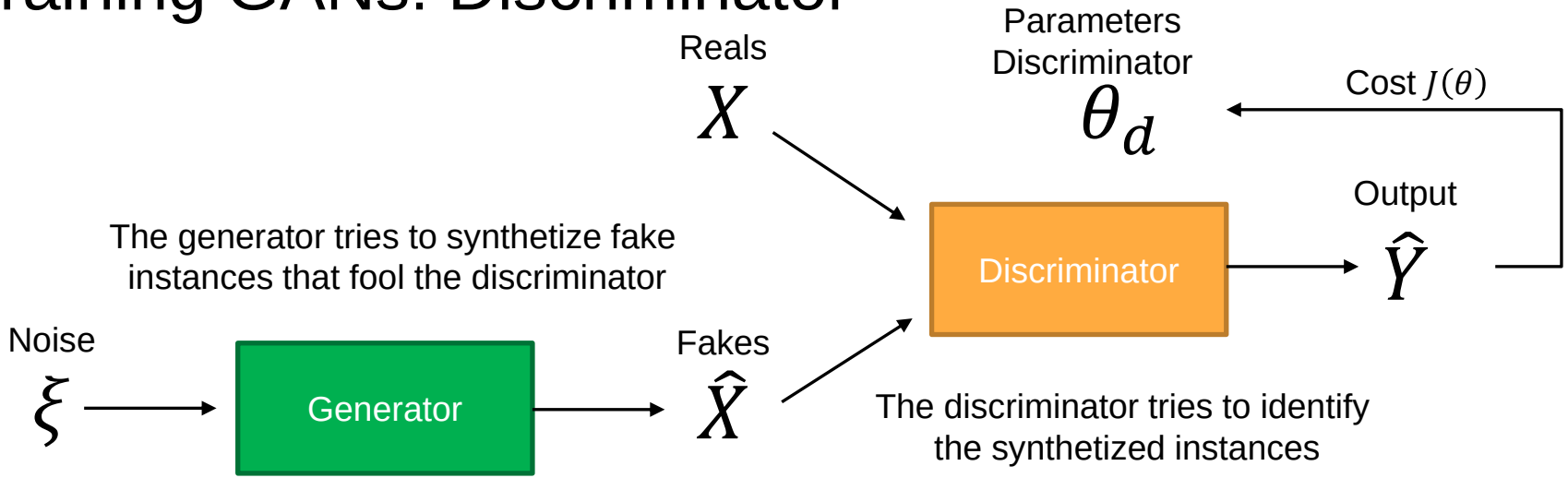
Goodfellow et al., “Generative adversarial nets”, NIPS’14

Training GANs



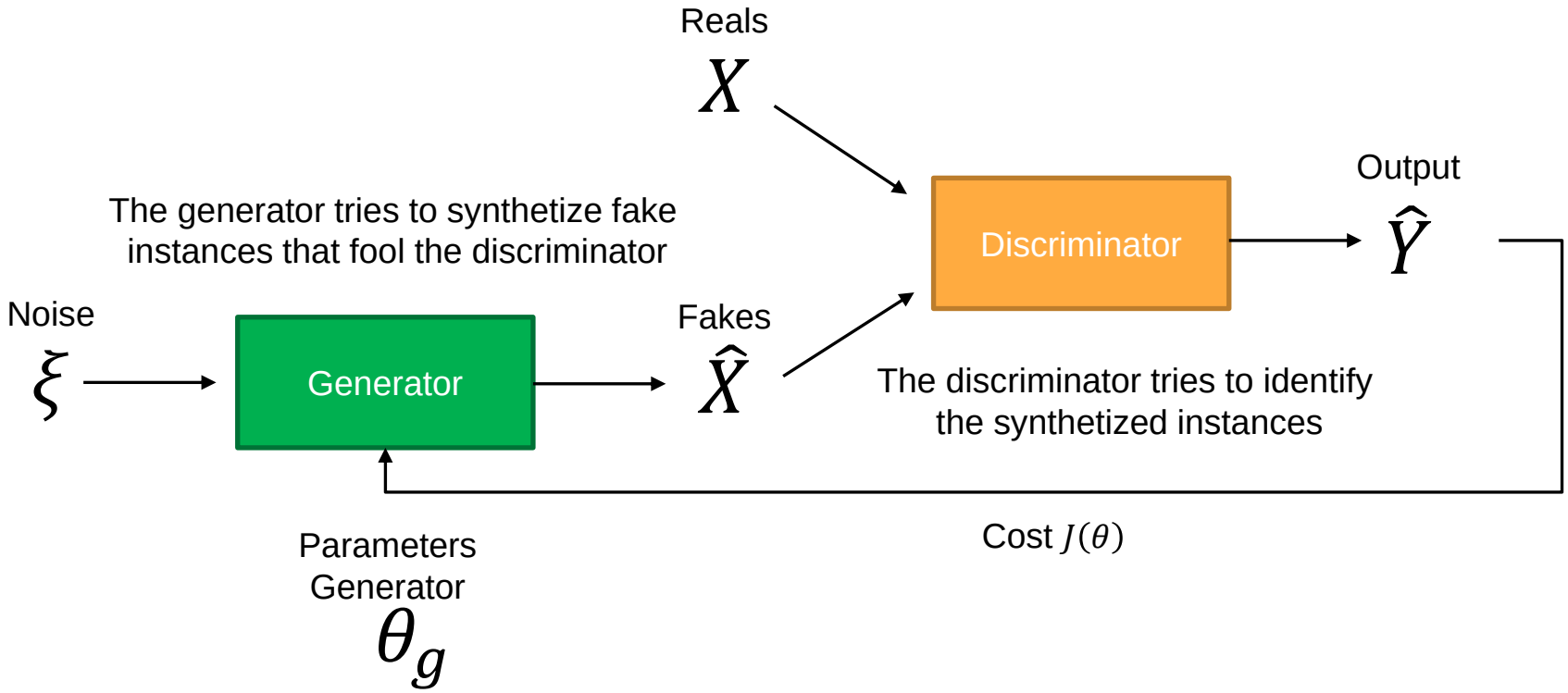
Training: adversarial objectives for the generator and the discriminator

Training GANs: Discriminator

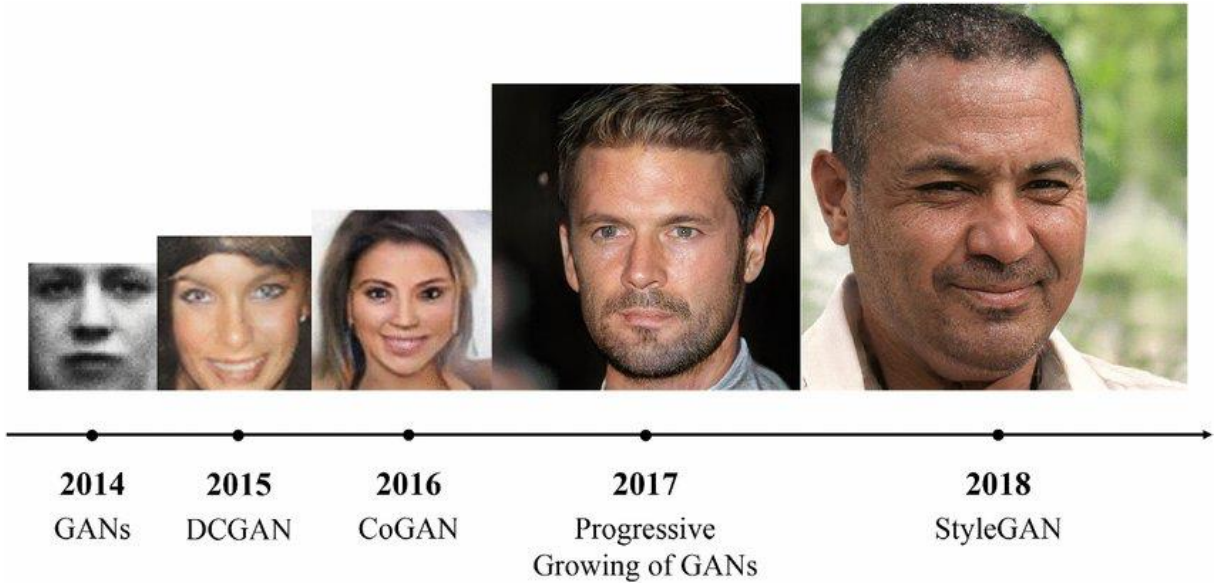


$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - \hat{y}^{(i)}\right)\right]$$

Training GANs: Generator



Growing of GANs



[Abdolahnejad et al., *Artificial Intelligence Review* 53.8 2020]

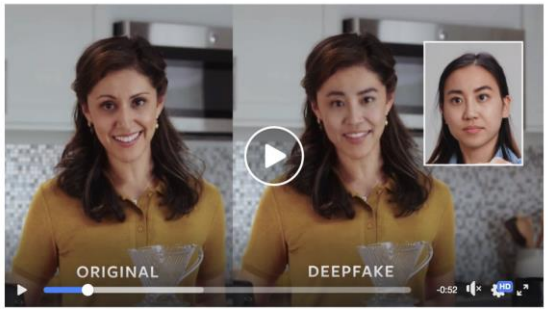
Deepfake Detection

Why is detection possible?

- Visual artifacts
- Semantic inconsistencies
- Camera-related artifacts
- Identity-related inconsistencies
- GAN fingerprints

A proliferation of datasets

- FaceForensics dataset: video dataset for forgery detection in human faces generated with the F2F facial reenactment algorithm altering facial expressions with the help of a reference actor
- FaceForensics++ (F2F, FaceSwap, DeepFake, Neural Textures) 1000 images for each manipulation methods
- Google and Jigsaw dataset
- Celeb-DF
- DeeperForensics-1.0
- Deepfake Detection Challenge Dataset (AWS, Facebook, Microsoft..)



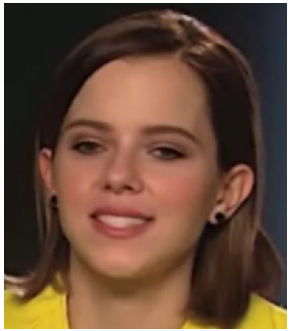
Dataset name	FaceForensics++			UADFV	Celeb-DF
	DeepFakes	Face2Face	FaceSwap		
Type	Identity swap	Expression reenactment	Identity swap	Identity swap	Identity swap
Generation	First	First	First	First	Second
DF quality	Low	High	Low	Low	High
Vis. artifacts	Yes	No	Yes	Yes	No


DeepFakes


Face2Face

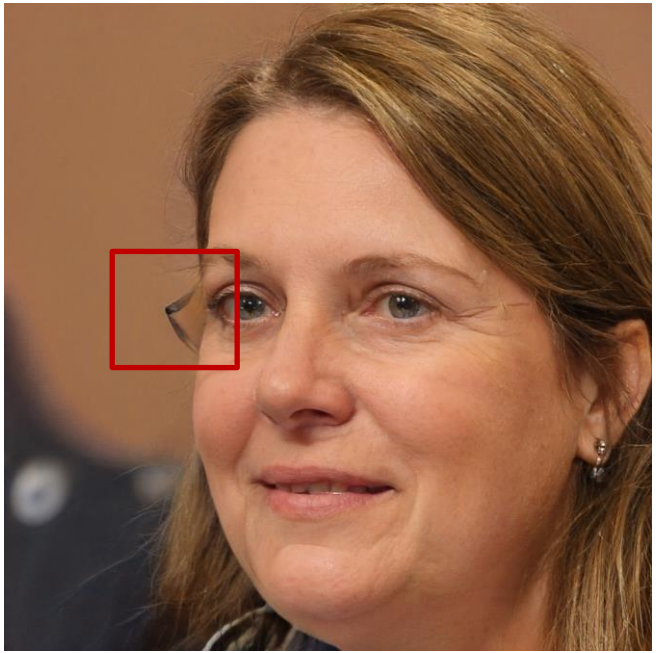

FaceSwap

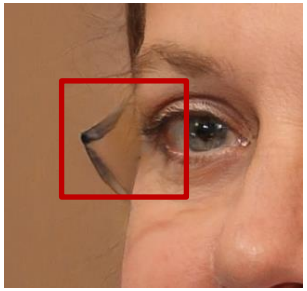

UADFV

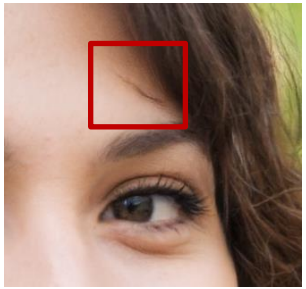

Celeb-DF

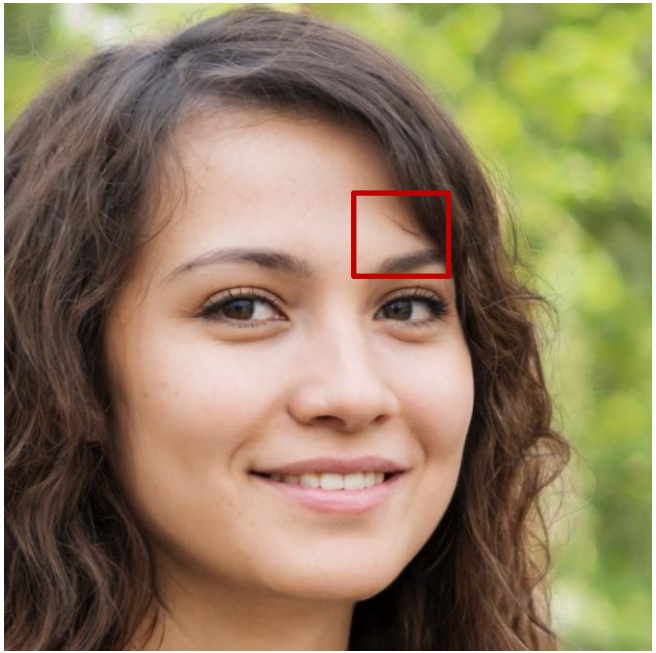
Visual Artifacts

Color anomalies, sharp boundaries, strange artifacts…



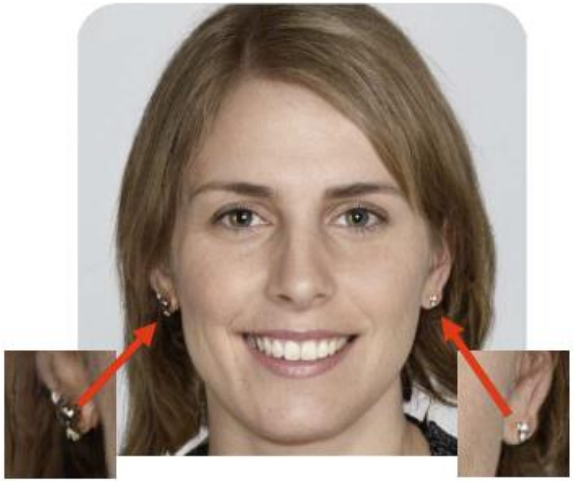






Semantic inconsistencies

Spatial inconsistencies in frames, semantic anomalies (e.g. different colour of the eyes), symmetry inconsistencies



Identity-related inconsistencies

Source/Target

Deepfake

Specific facial expressions and characteristics are not well preserved

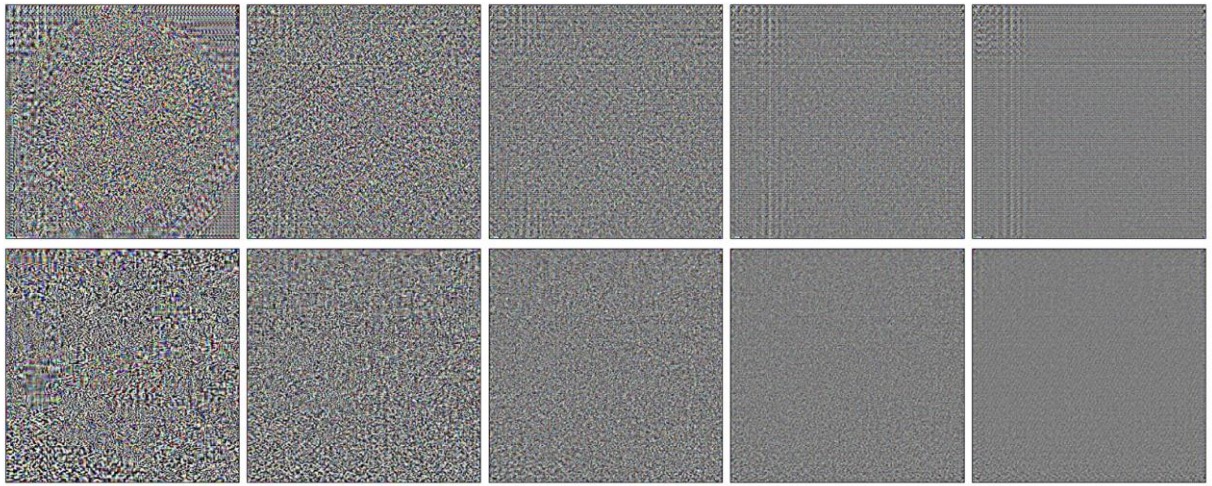
Google AI Blog: Contributing Data to Deepfake Detection Research (googleblog.com)

GAN fingerprints

GANs present specific artifacts due to the peculiar generation process

Cycle-GAN

Pro-GAN

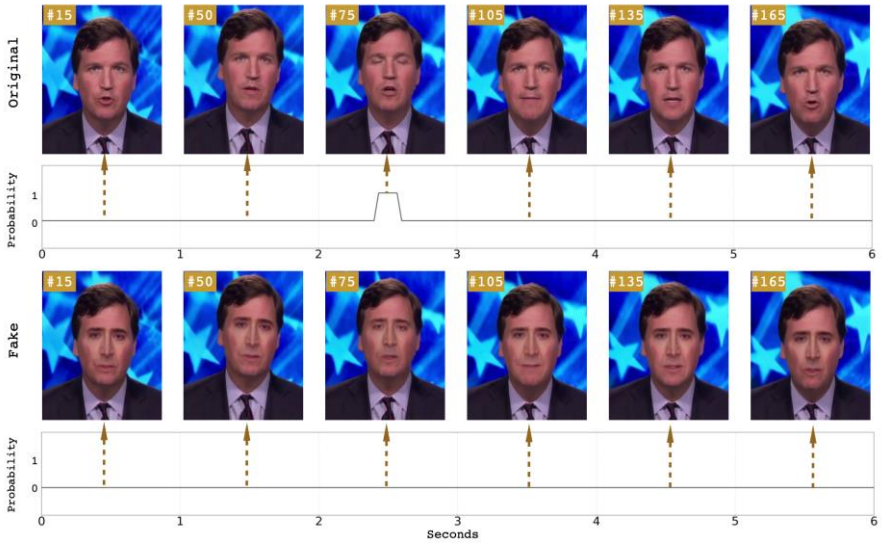


1. Marra et al. “Do GANs leave artificial fingerprints”, IEEE MIPR 2019
2. Yu et al., “Attributing Fake images to GANs: Learning and Analyzing GAN fingerprints”, ICCV 2019
3. Zhang et al., ”Detecting and simulating artifacts in GAN fake images”, IEEE WIFS 2019
4. Frank et al., “Leveraging Frequency Analysis for Deep Fake Image Recognition”, IEEE CVPR 2019

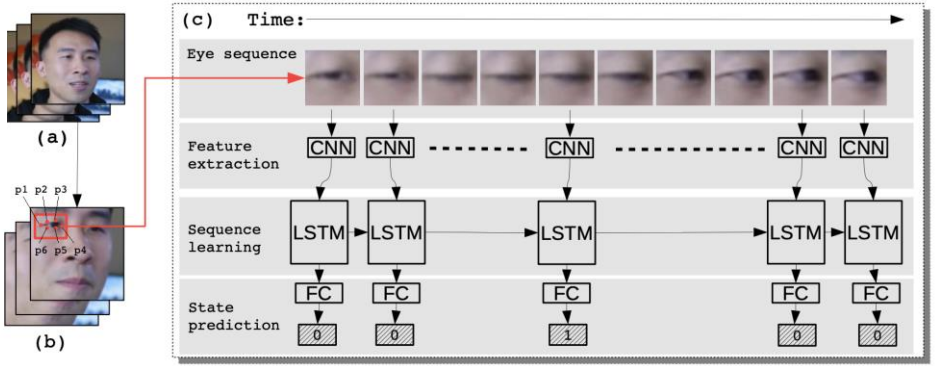
Detection strategies: Feature-based methods

- Eye blinking [Li18, Jung20]
- Corneal specular highlights [Hu20]
- Warping artifacts [Li19]
- Head pose inconsistencies [Yang19a]
- Landmark locations [Yang19b]
- Visual artifacts [Matern19]
- Heart variations [Fernandes19, Ciftci20, Hernandez-Ortega20, Qi20]
- Color cues [McCloskey18, Li18, Tondi20]
- Visual quality metrics [Korshunov18]
- Texture features [Bonomi20]

Eye blinking



In the original video (top), an eye blinking can be detected within 6 seconds, while in the fake video (bottom) such is not the case



Yuezun Li et al., “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking”, WIFS 2018

Heart variations

When blood moves through the veins, it changes the skin reflectance over time, due to the hemoglobin content in the blood. Photoplethysmography (PPG) signals can be extracted to recognize such changes by image processing techniques

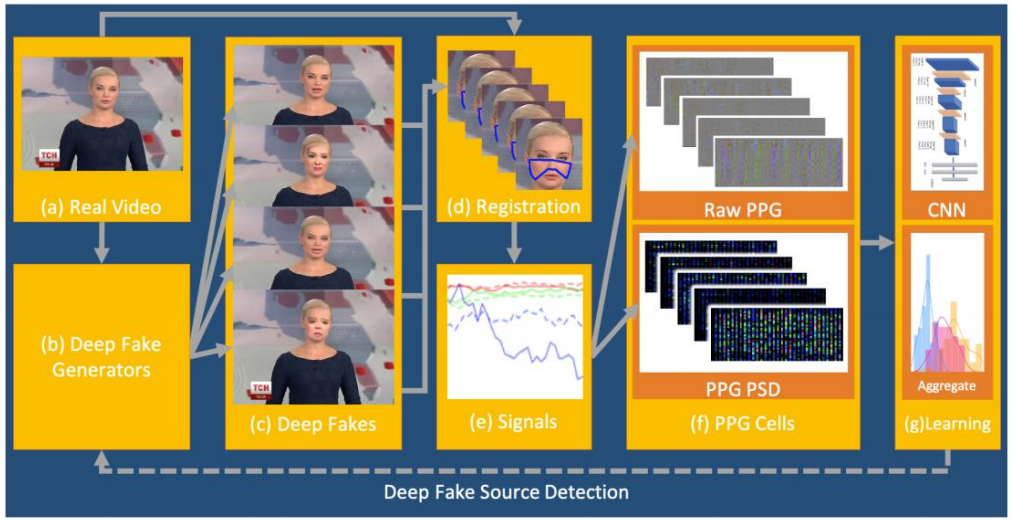
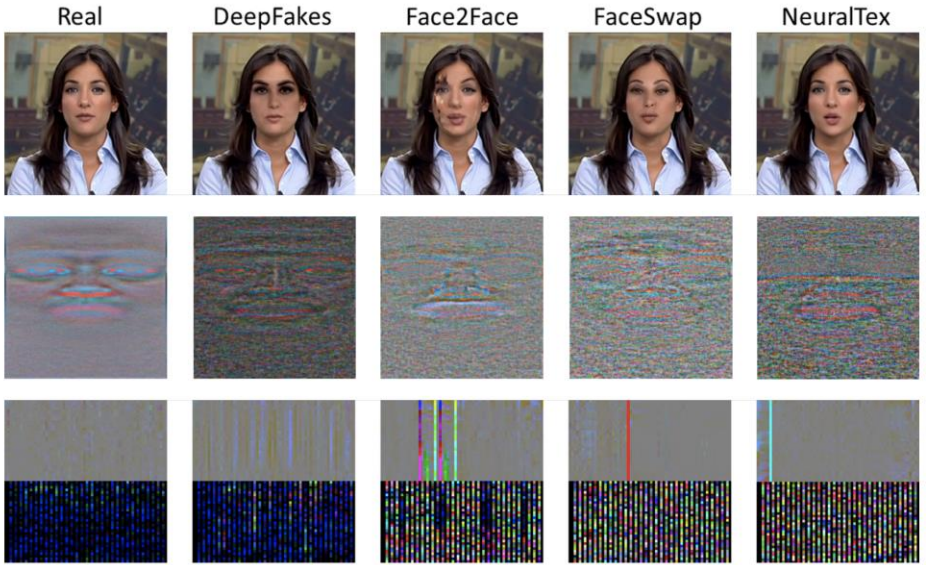


Figure 1. **Overview.** From real videos (a), several generators (b) create deep fakes with residuals specific to each model (c). Our system extracts face ROIs (d) and biological signals (e), to create PPG cells (f) where the residuals are reflected in spatial and frequency domains. Then it classifies both the authenticity and the source of any video (c) by training on PPG cells and aggregating window predictions (g).

Ciftci et al., "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals." IJCB 2020.

Heart variations

When blood moves through the veins, it changes the skin reflectance over time, due to the hemoglobin content in the blood. Photoplethysmography (PPG) signals can be extracted to recognize such changes by image processing techniques



Example frames per ω = 64 window (top), and their PPG cells (bottom) consisting of raw PPG and PPG PSD, of a real video (left) and its deep fakes per generative model (rest).

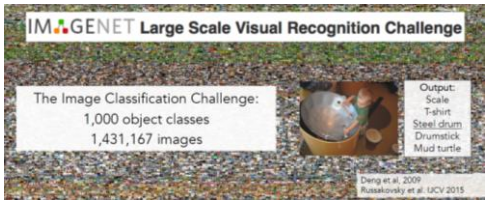
Ciftci et al., "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals." IJCB 2020.

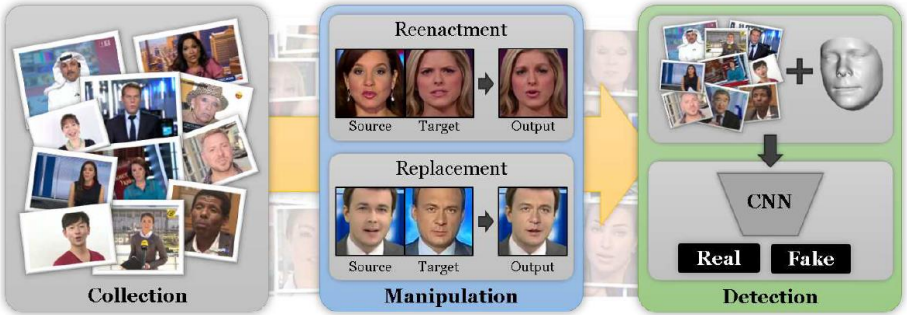
Detection strategies: Learning-based methods

- Pre-trained deep networks [Roessler19]
- MesoNet [Afchar18], CapsuleForensics [Nguyen19], Co-occurrenceNet [Nataraj19]
- Residual-based analysis [Cozzolino17, Guo20, Tariq20, Shinghal20]
- Recurrent networks [Guera18, Masi20, Montserrat20]
- Spatio-temporal features [Chen20, Ganiyusufoglu20, Wang20, Zhu20]
- Attention mechanisms [Dang20, Choi20, Mi20]
- Memory networks [Fernandes19]
- Fully convolutional networks [Tarasiou19]
- Frequency-based approaches [Zhang19, Durall20, Dzanic20, Qian20]
- Hybrid approaches [Chen20]
- GAN fingerprints [Marra19, Yu19]

Learning to Detect Manipulated Facial Images

- Face tracking method: extract the region of the image covered by the face; this region is fed into a learned classification network that outputs the prediction on the RGB patch
- Classification based on XceptionNet pretrained on ImageNet





[Rossler et Al, ICCV 2019]

The generalization issue

- In general, state-of-the-art Deepfakes detection methods are based on static frames features that though well-performing when trained on a specific kind of attack (same-forgery scenario), they show bad performances in a *cross-forgery scenario.*
- *Cross-forgery scenario:* when a model trained on a specific forgery is required to work against another unknown one.
- The generalization ability of forensics methods to other unseen types of generated fake content is taken into consideration so far especially focusing on GAN generated **images**.

Detection strategies: how to gain generalization

- Few-shot learning [Cozzolino18, Du19, Jeon19, Aneja20]
- Incremental learning [Marra19]
- Looking at common traces in fake faces [Li19]
- Patch-based analysis [Chai20]
- Augmentation [Xuan19, Wang20, Bondi20]
- Ensemble [Bonetti20, Rana20]
- One-class learning [Cozzolino19, Khalid20]
- Identity-based methods [Agarwal19, Agarwal20a, Agarwal20b, Cozzolino20]

Sequence-based approaches

Deepfake videos are usually detected by resorting at **frame-based** analysis

Sequence-based approaches by looking at possible dissimilarities in the video temporal structure

- Motion vectors should exploit different inter-frame correlations between fake and original videos and at certain extent generalization
- An approach based on **Optical-flow + CNN is proposed**

I. Amerini, R.Caldelli, “Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos”, 8th ACM Workshop on Information Hiding and Multimedia Security 2020.

SAPIENZA UNIVERSITÀ DI ROMA

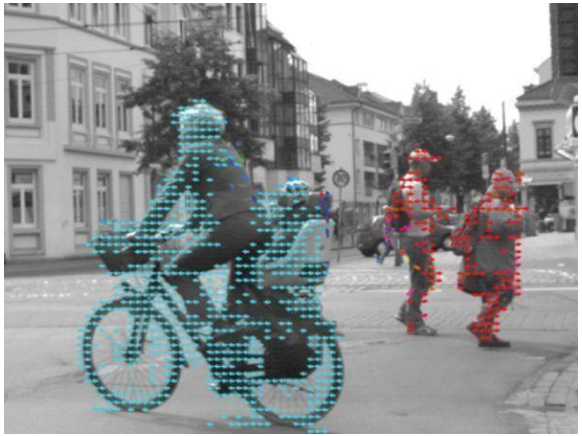
MICC

The optical flow approach

- *Optical flow fields* have been extracted from the video sequence
- Motion vectors should exploit different inter-frame correlations between fake and original videos
- Such an information is used as input of CNN-based classifiers

- **Optical Flow fields** describe the apparent motion of objects in a scene due to the relative motion between the observer (the camera) and the scene itself.
- Given two consecutive frames *f(t)* and *f(t+1)*:

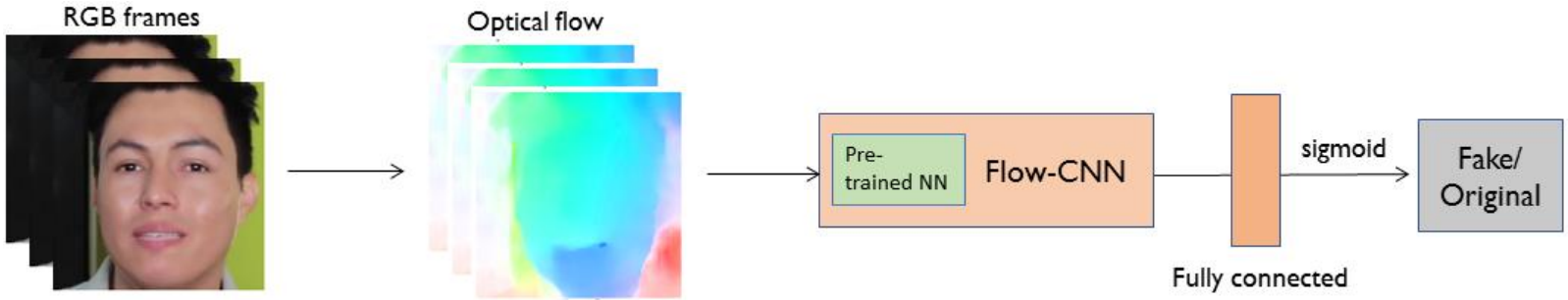
  $f(x, y, t) = f(x+\boldsymbol{\Delta x}, y+\boldsymbol{\Delta y}, t+1)$



R. Caldelli, L. Galteri, I. Amerini, A. Del Bimbo, “Optical Flow based CNN for detection of unlearnt deepfake manipulations, Pattern Recognition Letter 2021.
Amerini et Al, “Deepfake Video Detection through Optical Flow based CNN”, Human Behaviour and Understanding Workshop, ICCV 2019.

The proposed pipeline

- OF fields are used as input of a semi-trainable neural network
- Neural networks such as *ResNet50,* **pre-trained on Optical Flow**, have been tested
- The last convolutional layers and the dense ones are trained on deepfake dataset



Test set-up

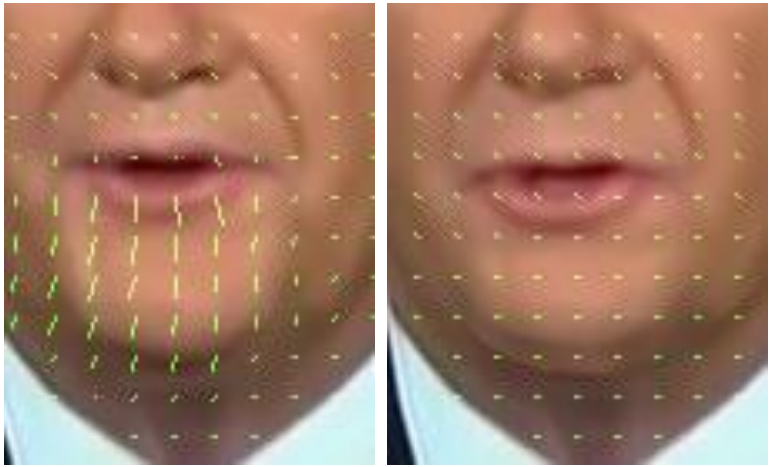
Dataset:

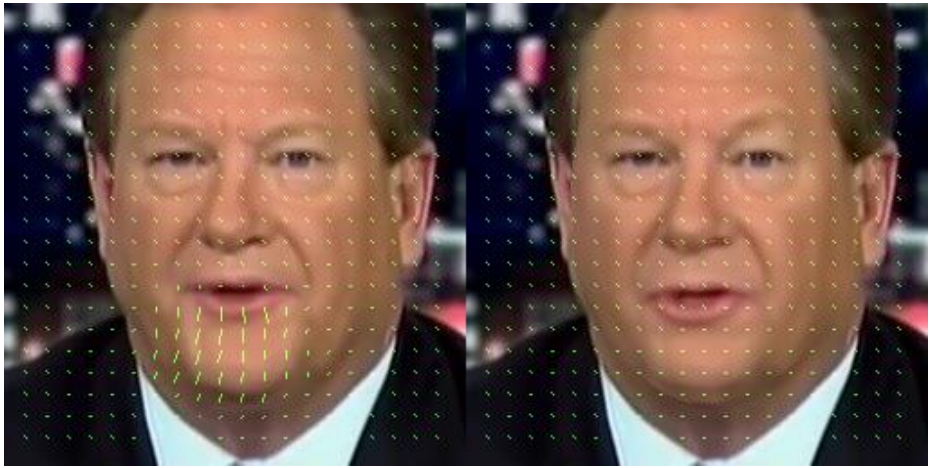
- *FaceForensics*++
- 1000 videos (original and fake for each kind of manipulation- Face2Face, Deepfake, Neural Texture)
- 720 for training set, 140 for validation and 140 for test set

- A patch of 300x300 pixels, around the face, is cropped from each frame
- A squared patch of 224x224 pixels is randomly chosen and flipped left-right for data augmentation
- Adam optimizer with learning rate 10^{-4}, default momentum values and batch size of 256 is used

Experimental results – qualitative evaluation

- Looking at MVs, particularly around the mouth, a different distribution of the OF field is appreciable:
    - Deepfake case is smoother



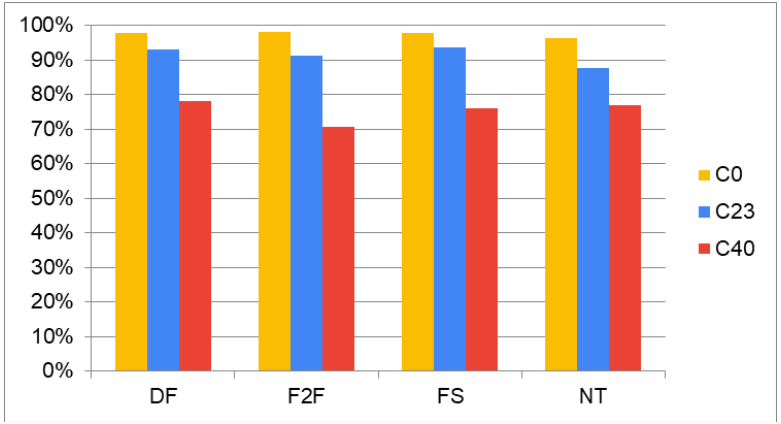


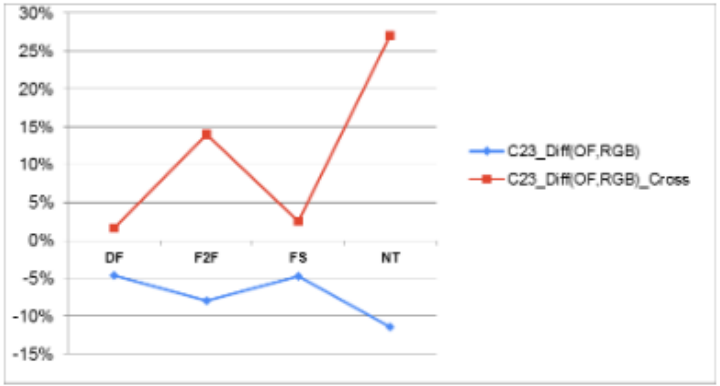
REAL DEEPFAKE

Experimental results

- Accuracy **higher than 90%** for *FaceForensics*++ dataset (Face2Face, DeepFake, FaceSwap, NT).

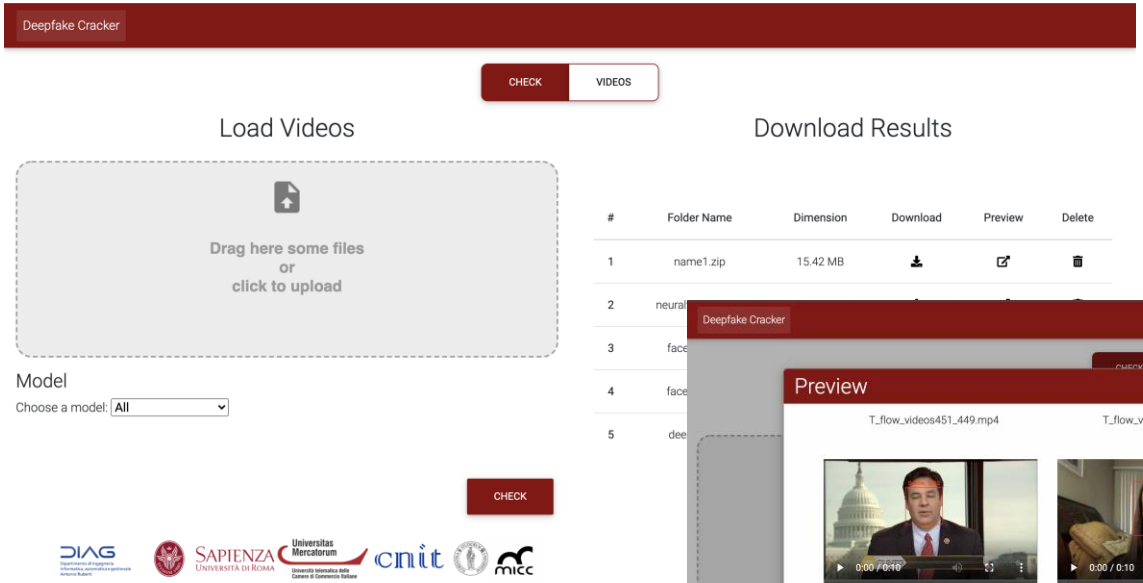
- This kind of feature is suited to extract peculiar features between the fake and real cases, especially when working in the challenging *cross-forgery scenario*

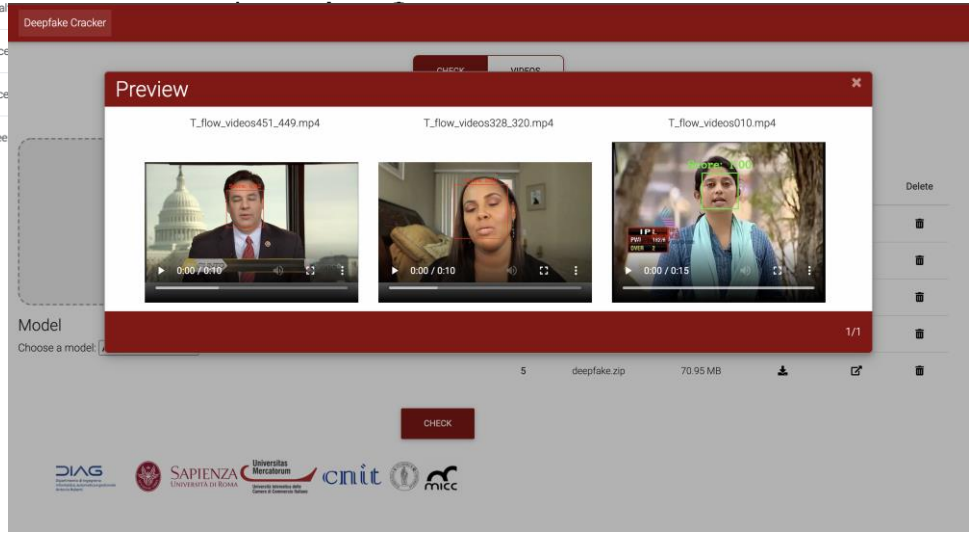




DeepFake Cracker tool

https://www.youtube.com/watch?v=tfIuA7cfmEk&t=25s





Demo presented at the International Conference on Pattern Recognition ICPR 2021

Future trends: media forensics in the wild

- The ease of dissemination of fake content through social media platforms makes the ability to reconstruct the source of images and videos increasingly important

- Source identification on shared data
  - Both device identification and social network provenance need to be examined in depth
- Forgery detection on shared data
  - Deepfake on social media not only on lab datasets
  - It becomes useful to take into consideration multi-modal media assets like the text associated with the image or video as well, which can be useful to improve the semantic analysis of fake content.
- Tool for verifying machine learning/deep learning models

Questions?

- Drop us an email amerini@diag.uniroma1.it

  maiano@diag.uniroma1.it

- https://sites.google.com/diag.uniroma1.it/ireneamerini

- Alcor Lab Via Ariosto 25, Rome

Multimedia Forensics
Green AI (Vertical Farming and Beekeeping)
Edge-Vision
Deep Learning Theory
Visual Knowledge acquisition: Activity Recognition & Object Detection

Acknowledgement: some slides and material from Sharon Zhou, Matthias Niesser, Luisa Verdoliva and Ava Soleimany

Media Forensics and the challenge of Deepfakes
in a social media context

Irene Amerini
amerini@diag.uniroma1.it

Luca Maiano
maiano@diag.uniroma1.it

19/05/2021

SAPIENZA
UNIVERSITÀ DI ROMA