DeepFake Detection: state-of-the-art and future directions

Luisa Verdoliva
Università degli Studi di Napoli Federico II

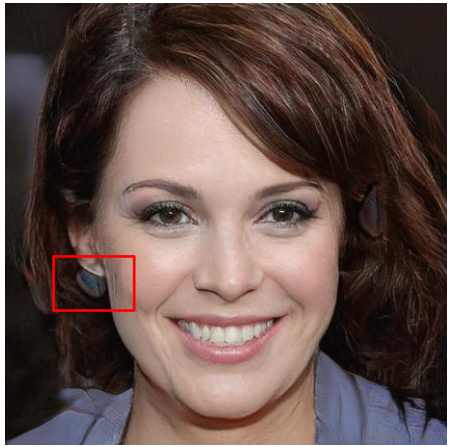
2nd JPEG Fake Media Workshop
March 25, 2021

Why is detection possible?

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

Visual artifacts

- Color anomalies



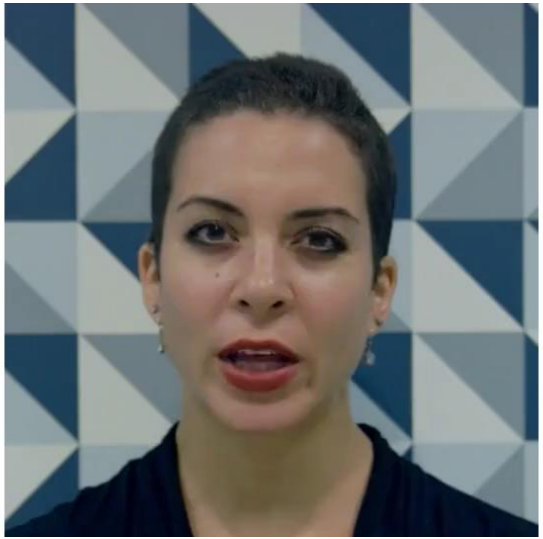
Semantic inconsistencies

- Lack of symmetry (e.g. different eye color, ears, earrings)

Identity related inconsistencies

- The specific face expression of the source identity are not well preserved

Source Identity



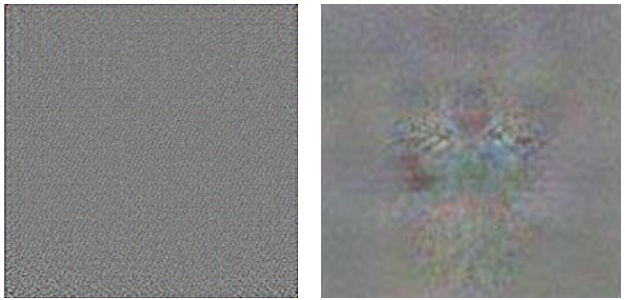
Target Video

Deepfakes

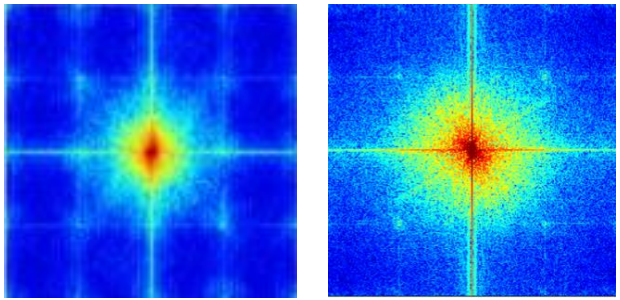
GAN-specific traces

- Synthetic images generated by a GAN present specific artifacts because of the peculiar generation process

Artificial fingerprints [1,2]



Frequency domain traces [3,4]



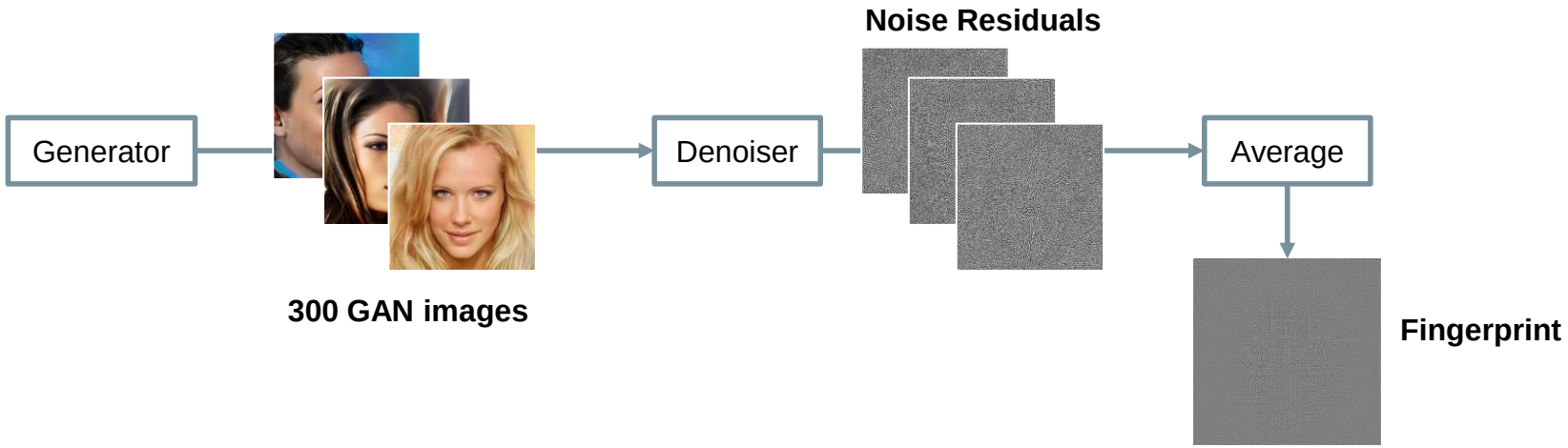
[1] Marra et al., “Do GANs leave artificial fingerprints”, *IEEE MIPR* 2019.

[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.

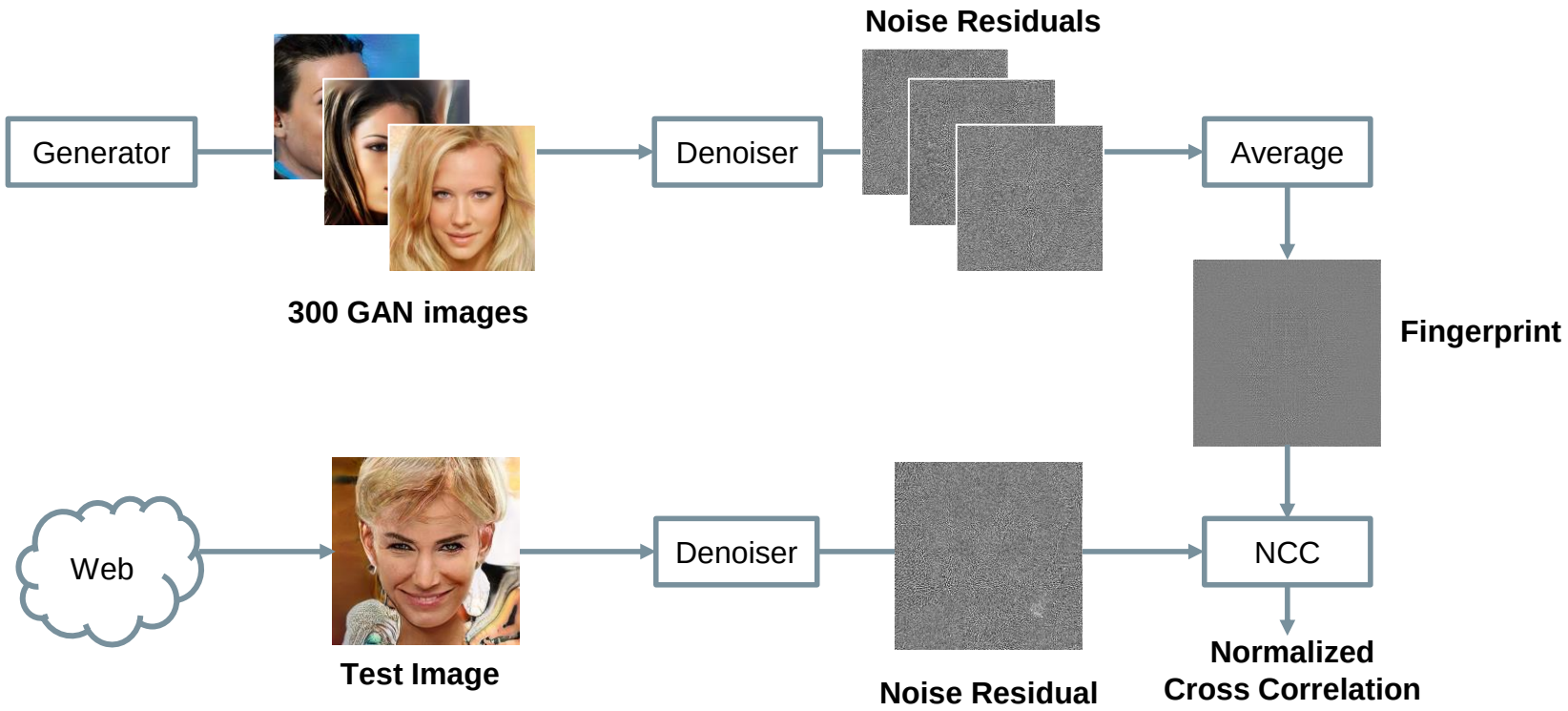
[4] Frank et al., “Leveraging Frequency Analysis for Deep Fake Image Recognition”, *IEEE CVPR* 2020.

PRNU-like procedure



Marra et al., “Do GANs leave artificial fingerprints?” *IEEE Workshop on Fake Multimedia*, March 2019.

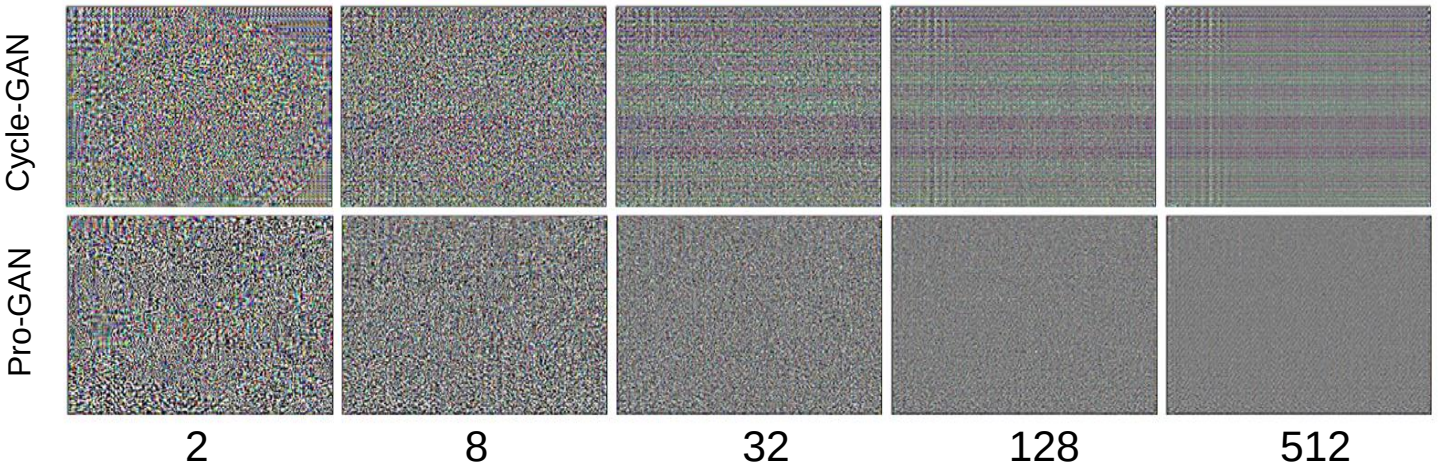
PRNU-like procedure



Marra et al., “Do GANs leave artificial fingerprints?” *IEEE Workshop on Fake Multimedia*, March 2019.

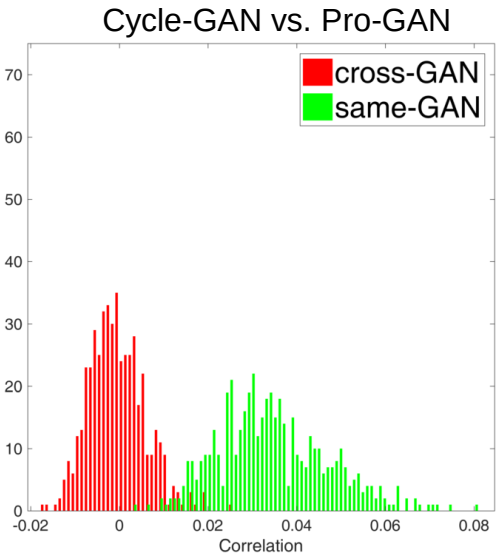
GAN fingerprints

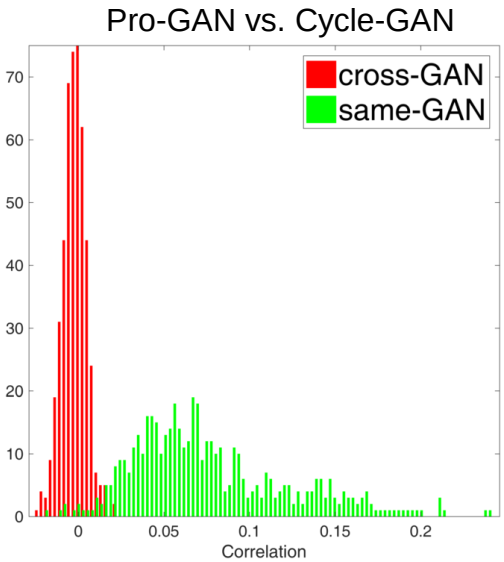
- Fingerprints of two GANs, estimated over a growing number of residuals



Comparing GAN fingerprints

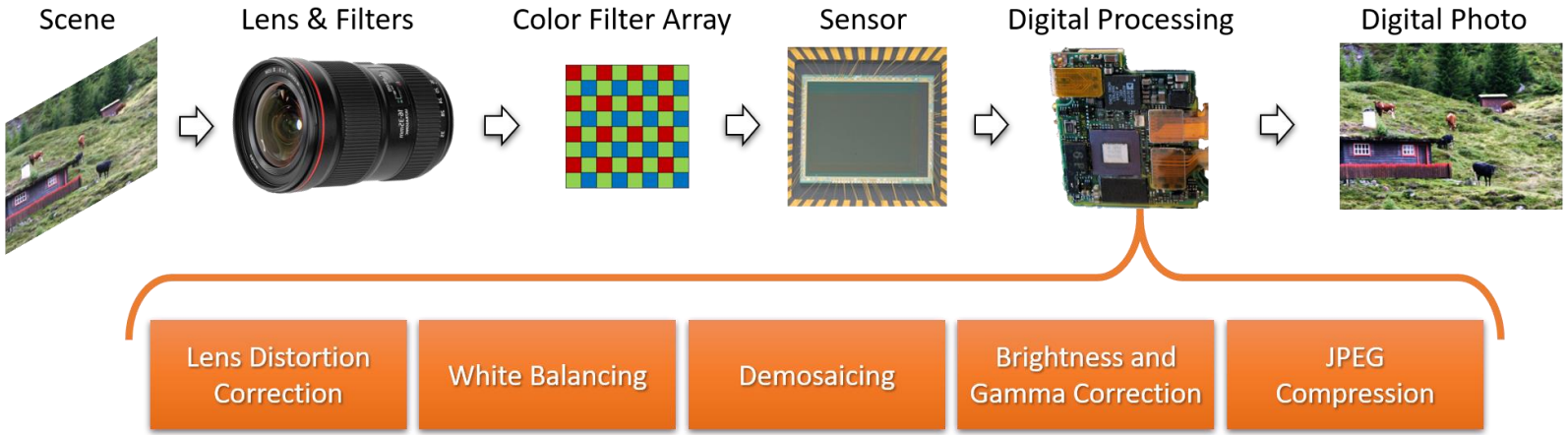
- Cross-GAN (red) and Same-GAN (green) correlations are well separated indicating the presence of a unique fingerprint





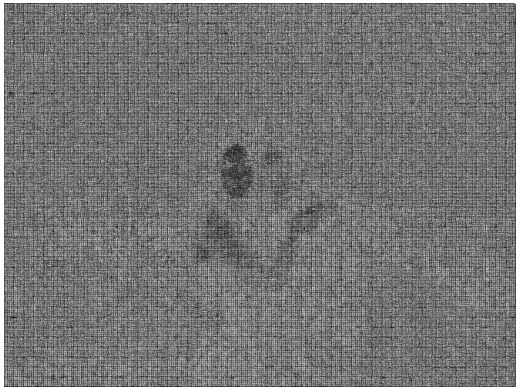
Camera-related artifacts

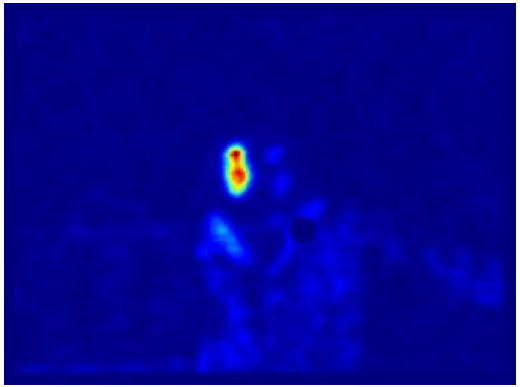
- In-camera operations



Manipulations as anomalies

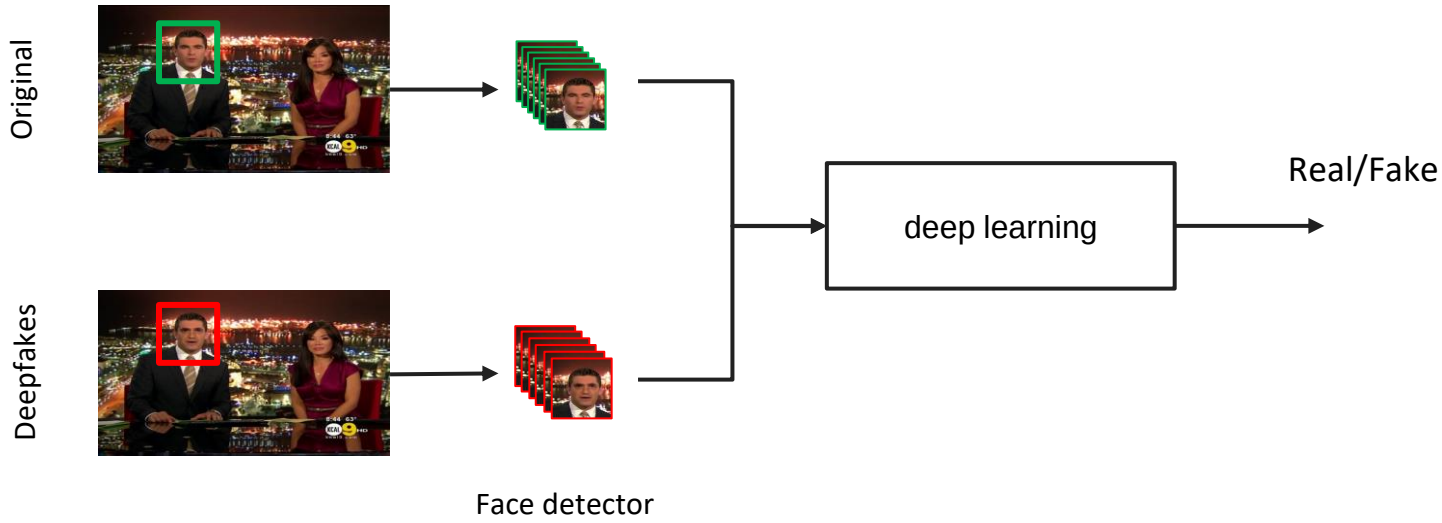
- It is possible to highlight these traces by extracting a camera fingerprint





Deepfake detection: supervised learning

- Train on large datasets of pristine and fake videos to learn the artifacts (visible and not visible)



Learning-based methods

- MesoNet [Afchar18], CapsuleForensics [Nguyen19], Co-occurenceNet [Nataraj19]
- Pre-trained deep networks [Roessler19]
- Residual-based analysis [Cozzolino17, Guo20, Tariq20, Singhal20]
- 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]

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]

FaceForensics++

1000 original videos + manipulated videos using

- FaceSwap
- Face2Face
- DeepFake
- Neural Textures

+ 3000 manipulated videos from **Google AI**

A. Roessler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, “FaceForensics++: Learning to detect manipulated facial images”, ICCV 2019.

FaceForensics++: example

Target Video

Source Video

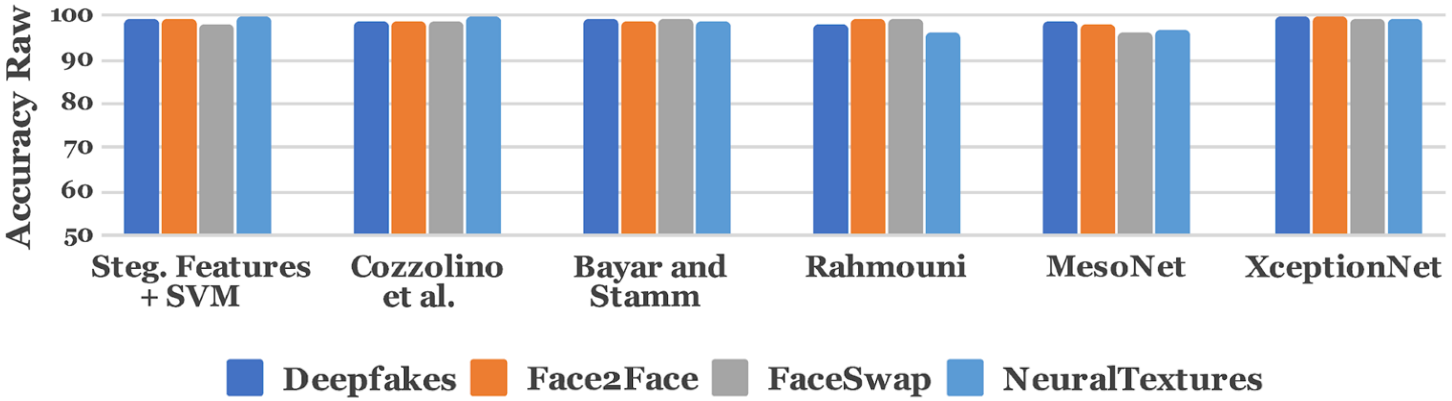
FaceSwap

Face2Face

DeepFake

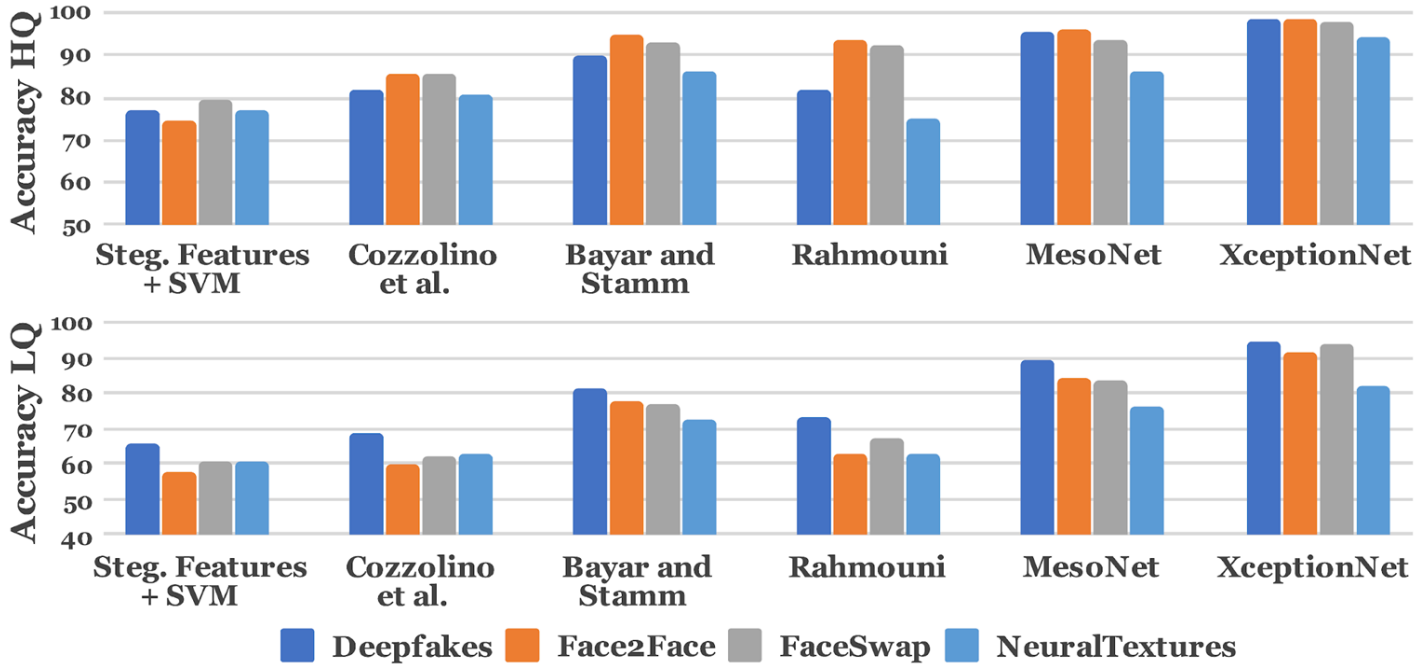
Neural Textures

Detection results (uncompressed data)



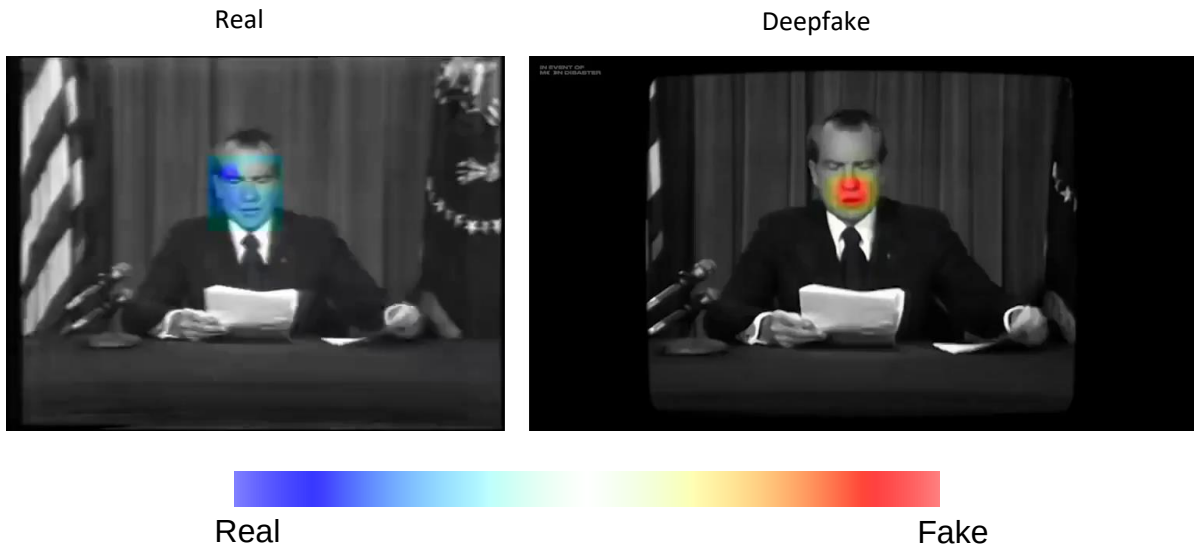
Fridrich and Kodovsky, “Rich Models for Steganalysis of Digital Images,” *IEEE TIFS* 2012
B.Bayar and M.Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer”, *ACM IH&MMSec* 2016
Cozzolino et al., “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection”, *ACM IH&MMSec* 2017
Rahmouni et al., “Distinguishing computer graphics from natural images using convolution neural networks” *IEEE WIFS* 2017
Afchar et al., “MesoNet: a compact facial video forgery detection network”, *IEEE WIFS* 2018
Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions”, *IEEE CVPR* 2017

Detection results (compressed data)



Interpretability: CAM visualization

Deepfake video created by the MIT Center for Advanced Virtuality (https://virtuality.mit.edu/)



More recent large deepfake datasets

- Celeb-DF (2020): 590 pristine – 5,639 forged
- Facebook dataset DFDC (2020): 19,154 pristine – 100,000 forged
- DeeperForensics (2020): 50,000 pristine – 10,000 forged
- WildDeepfake (2020): 3,800 pristine – 3,500 forged

Li et al., “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,” *IEEE CVPR* 2020
Dolhansky et al., “The DeepFake Detection Challenge Dataset”, arXiv:2006.07397v3, 2020
Jiang et al., “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection”, *IEEE CVPR* 2020
Zi et al., “WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection”, *ACM Multimedia* 2020

Cross-dataset analysis

- We can conduct a cross-dataset analysis to check for the generalization ability of the CNN models (FF++ vs DFDC)

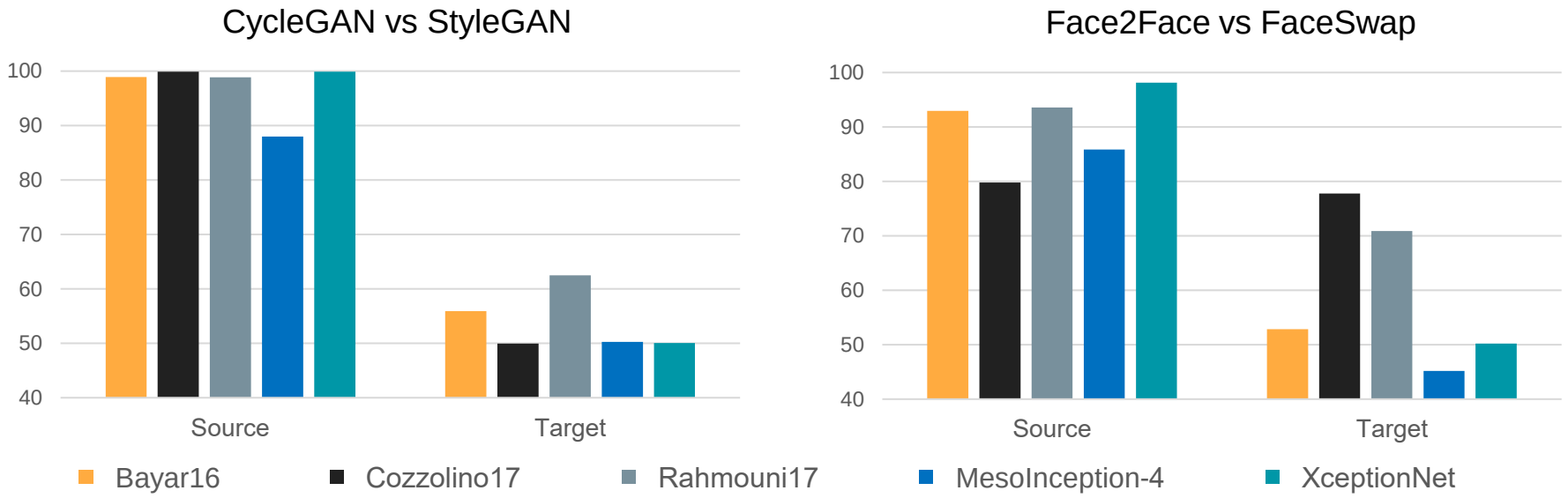
XceptionNet	Test on FF++	Test on DFDC
Train on FF++	95.52%	62.63%
Train on DFDC	70.14%	91.90%

Cross-dataset analysis

- We can conduct a cross-dataset analysis to check for the generalization ability of the CNN models (Face2Face vs FaceSwap)

XceptionNet	Test on Face2Face	Test on FaceSwap
Train on Face2Face	98.13%	50.20%
Train on FaceSwap	51.73%	98.30%

More results on generalization



B.Bayar and M.Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer”, *ACM IH&MMSec* 2016

Cozzolino et al., “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, *ACM IH&MMSec* 2017

Rahmouni et al., “Distinguishing computer graphics from natural images using convolution neural networks” *IEEE WIFS* 2017

Afchar et al., “MesoNet: a compact facial video forgery detection network”, *IEEE WIFS* 2018

Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions”, *IEEE CVPR* 2017

CycleGAN vs StyleGAN

- Different architectures to perform image-to-image translation

Real

CycleGAN

Real

StyleGAN

Zhu et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *ICCV 2019*
Karras et al., “A style-based generator architecture for generative adversarial networks,” *CVPR* 2019

Considerations

- In a supervised setting deep learning approaches perform very well, but…
- only if the training includes data with the target manipulation
- This holds both for different facial manipulations and for GAN synthetic generated images

How to gain generalization

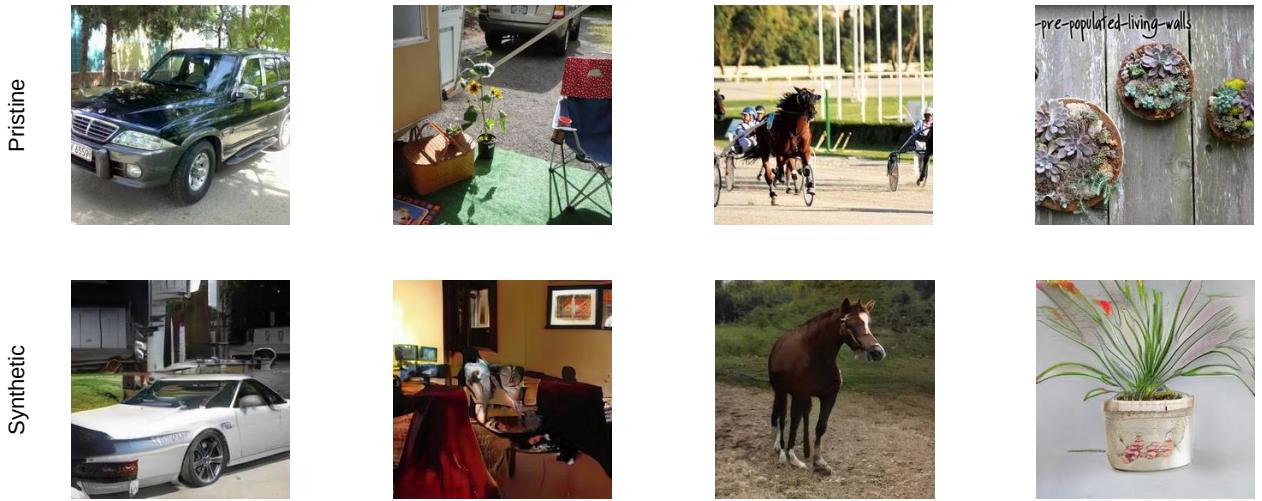
- Few-shot learning [Cozzolino18, Du19, Jeon19, Aneja2020]
- Incremental learning [Marra19]
- Looking at common traces in fake faces [Li19]
- Patch-based analysis [Chai20]
- Augmentation [Xuan19, Wang20, Bondi20]
- Ensemble [Bonettini20, Rana20]
- One-class learning [Cozzolino19, Khalid20]
- Identity-based methods [Agarwal19, Agarwal20a, Agarwal20b, Cozzolino20]

How to gain generalization

- Few-shot learning [Cozzolino18, Du19, Jeon19, Aneja2020]
- Incremental learning [Marra19]
- Looking at common traces in fake faces [Li19]
- Patch-based analysis [Chai20]
- Augmentation [Xuan19, Wang20, Bondi20, Gragnaniello21]
- Ensemble [Bonettini20, Rana20]
- One-class learning [Cozzolino19, Khalid20]
- Identity-based methods [Agarwal19, Agarwal20a, Agarwal20b, Cozzolino20]

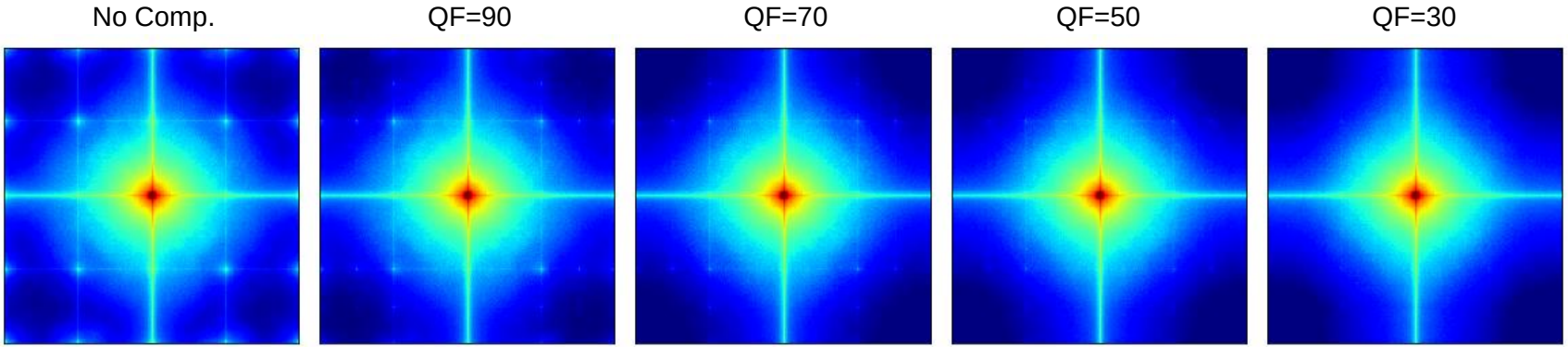
Augmentation

- Training using only one GAN architecture: ProGAN (LSUN) and strong augmentation (standard operations + blurring + JPEG compression)

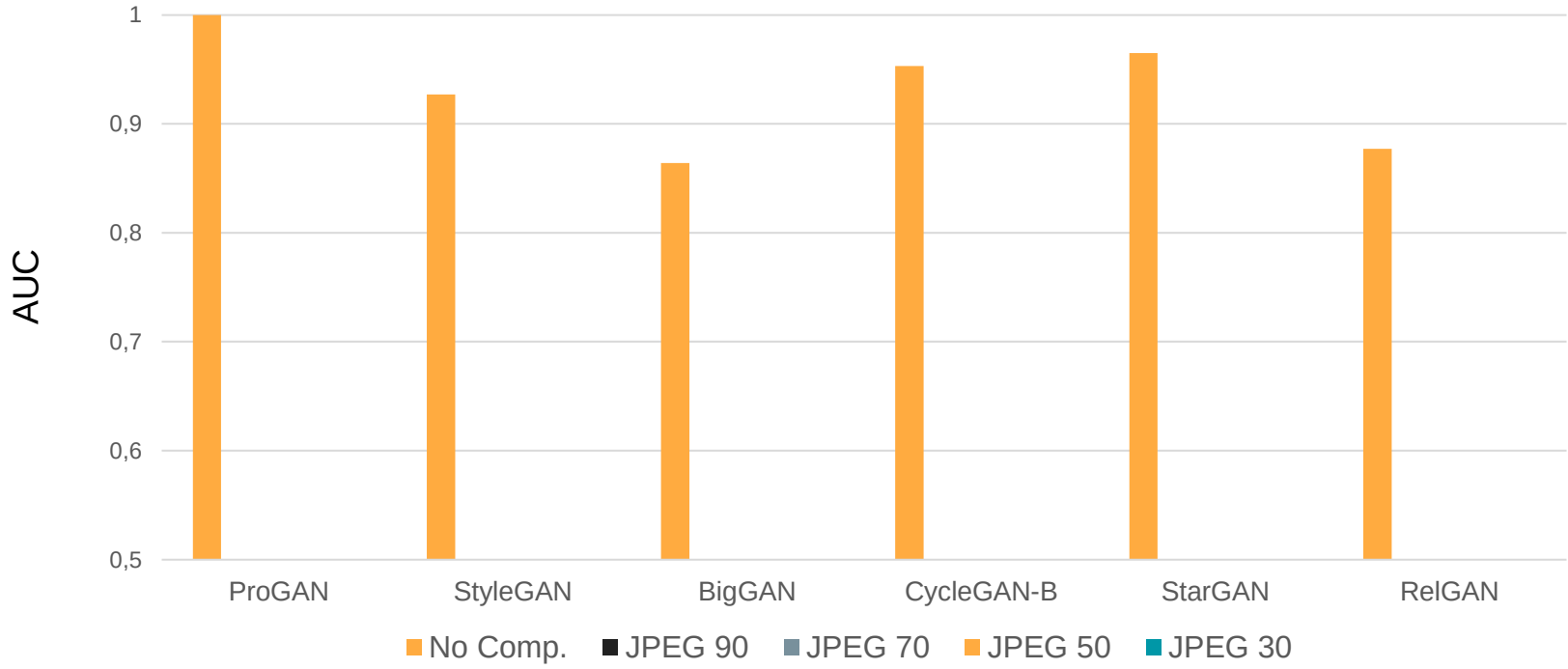


Wang et al., “CNN-generated images are surprisingly easy to spot... for now”, *IEEE CVPR 2020.*

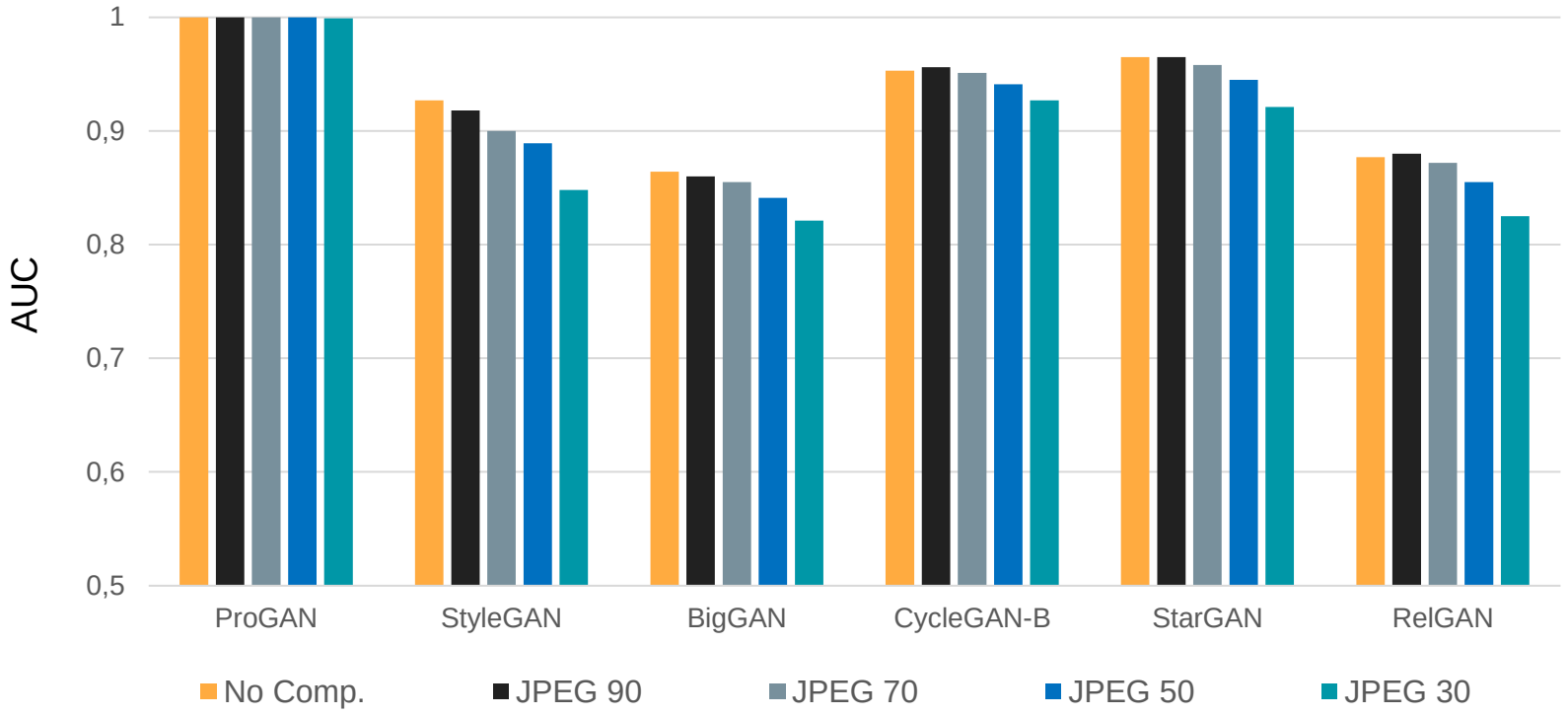
GAN traces tend to vanish after compression



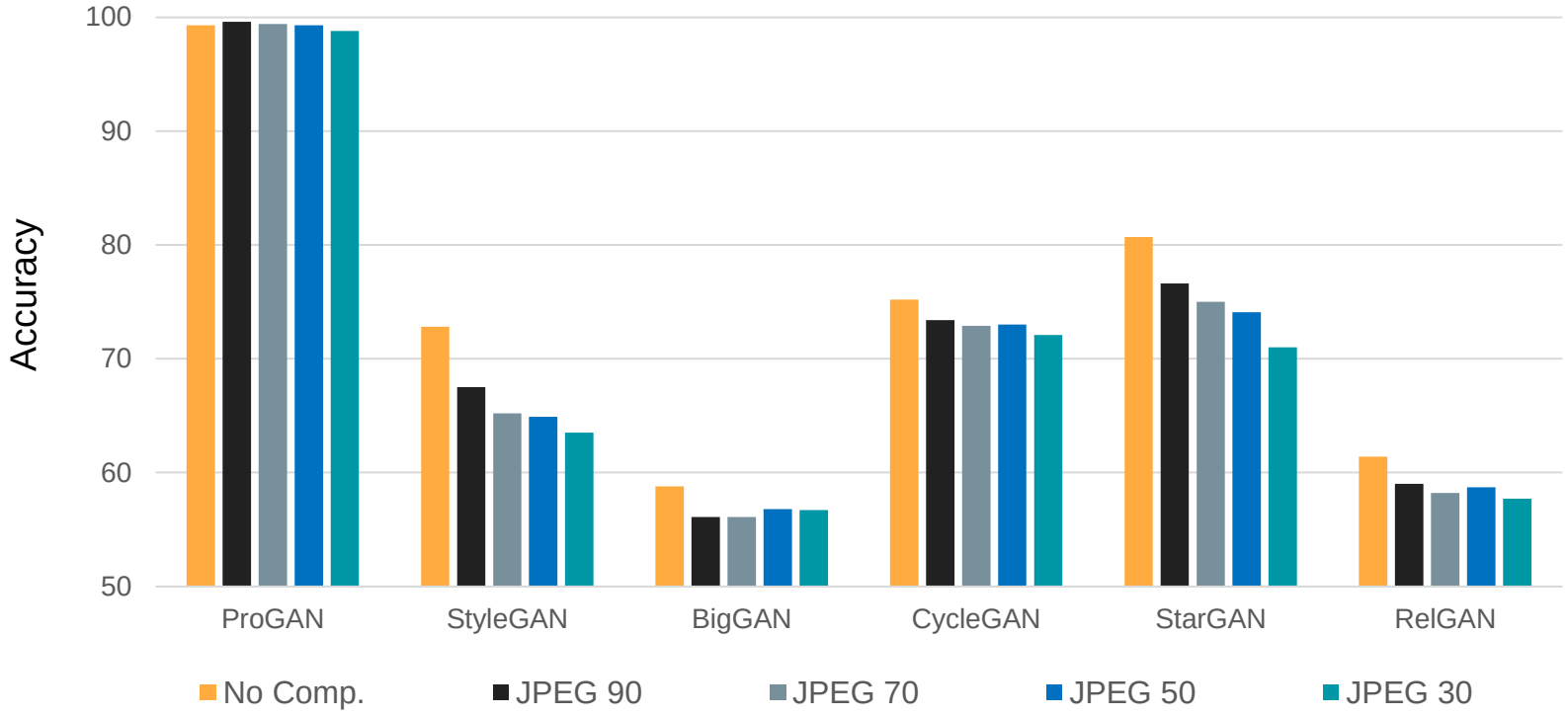
GAN Detection on unseen architectures



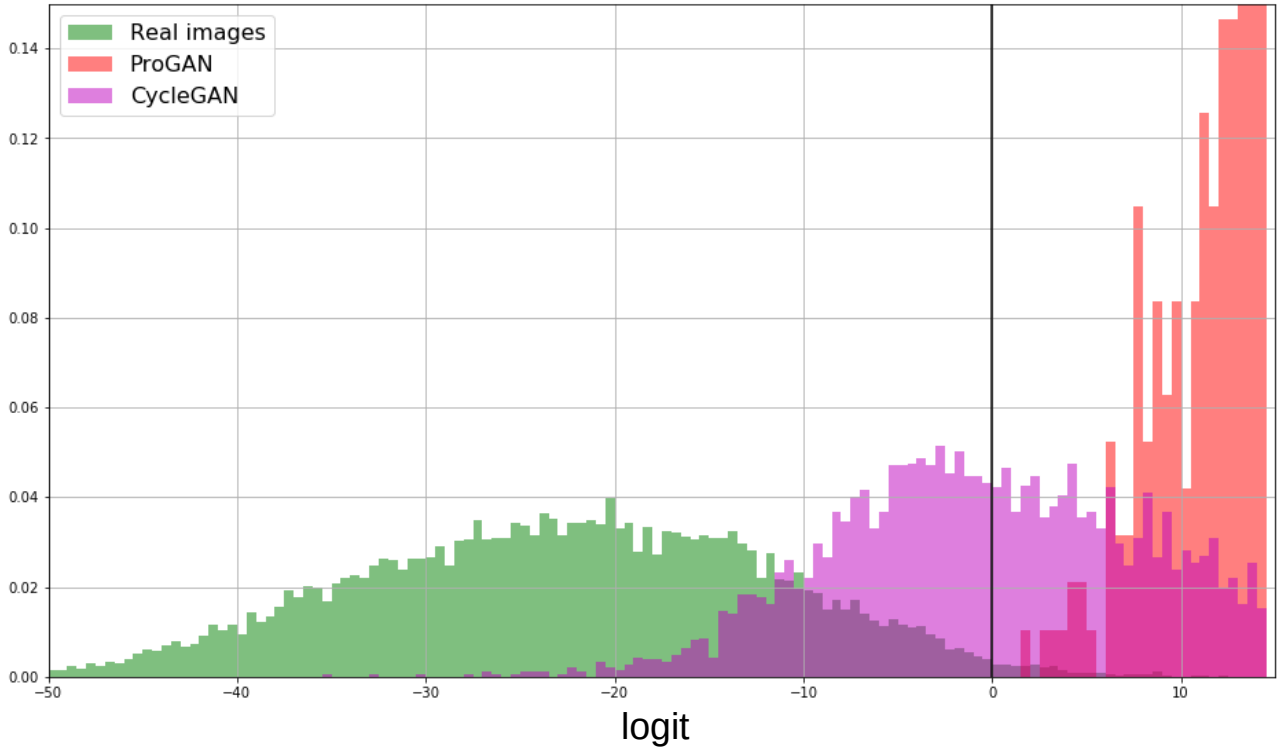
Varying compression level (AUC)



Varying compression level (Accuracy)



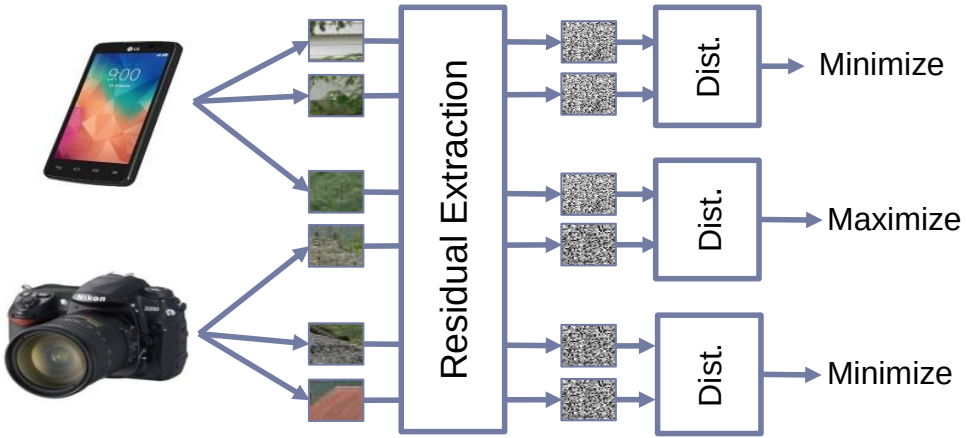
Threshold sensitivity



Gragnaniello et al., “Are GAN generated images easy to detect? A critical analysis of the state-of-the-art”, *IEEE ICME 2021.*

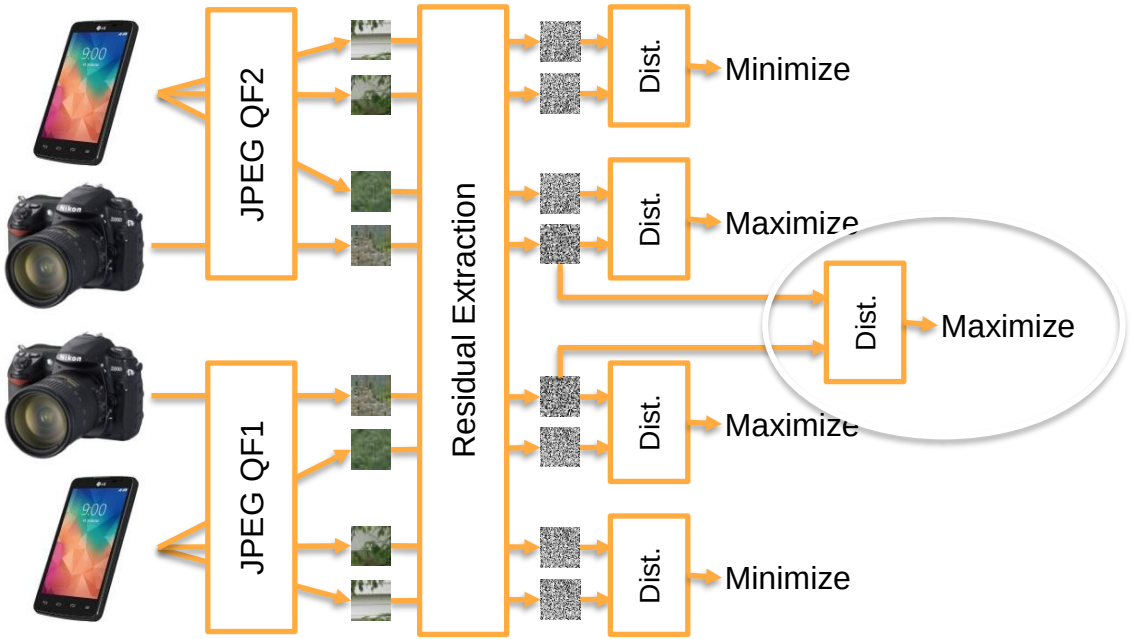
One-class learning (1)

- We train a siamese network to:
  - minimize the distance between residual patches from the *same camera and position*
  - maximize the one from different cameras



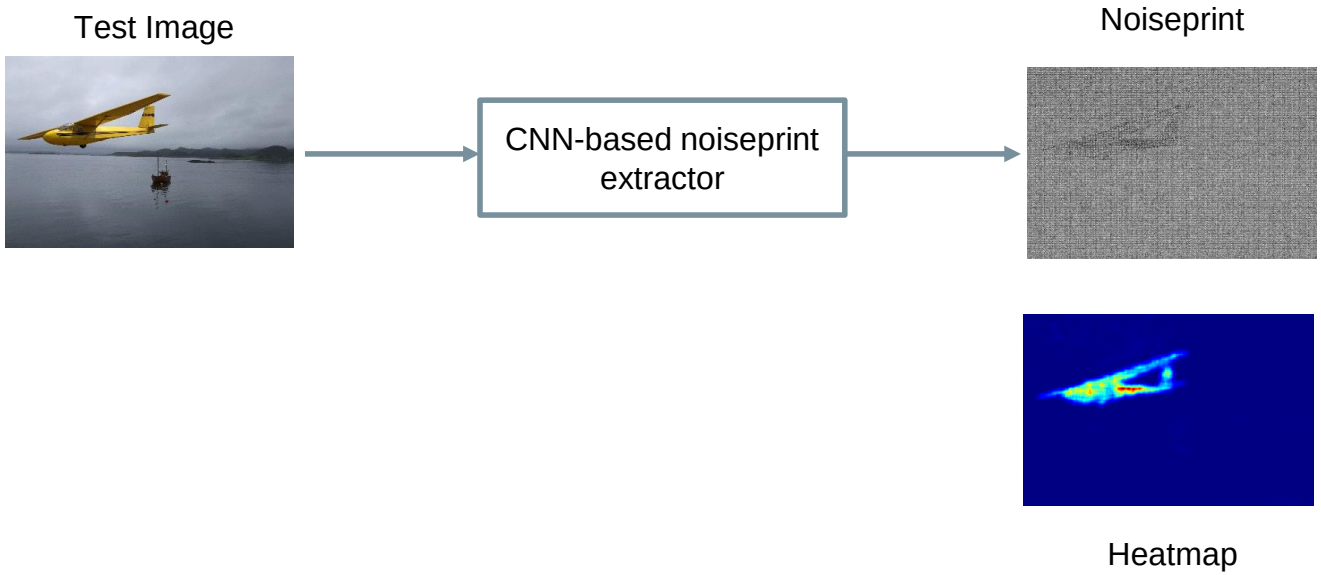
One-class learning (2)

- We enhanced this procedure by also including the JPEG history:
  - patches at different compression levels are compared and considered negative couples



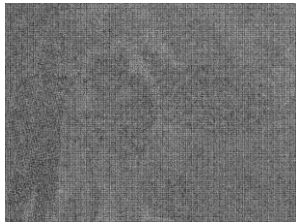
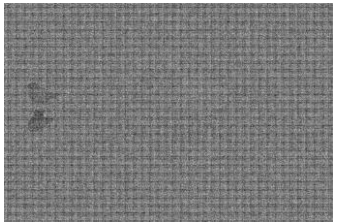
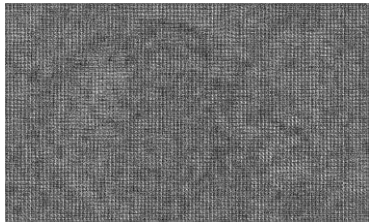
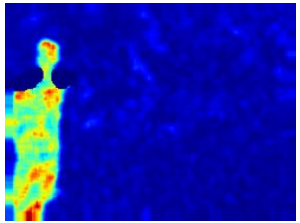
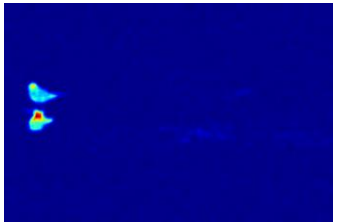
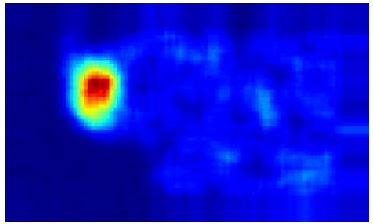
Noiseprint

- The extracted noise residual (noiseprint) can enhance traces coming from different cameras or editing-based anomalies



Cozzolino and Verdoliva, “Noiseprint: a CNN-based camera model fingerprint,” *IEEE TIFS*, 2019.

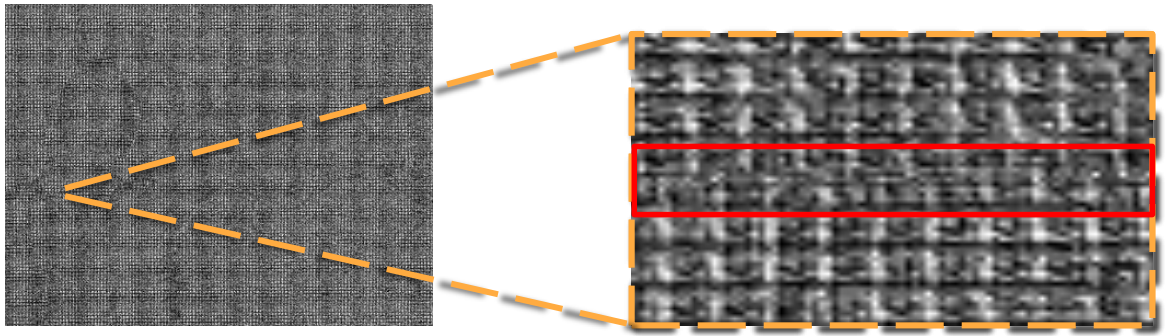
Sample results

	Splicing	Copy-Move	Faceswap
Test Image			
Noise Residual			
Heat map			

Some more insights

- JPEG grid misalignment

Image



Noiseprint

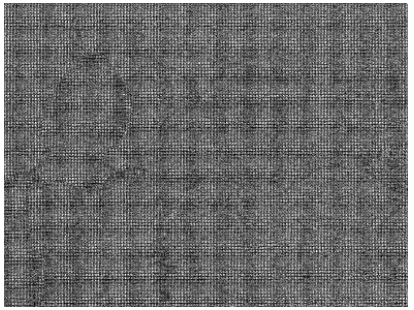
Zoom

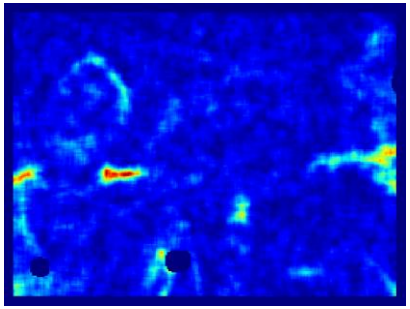
Some more insights

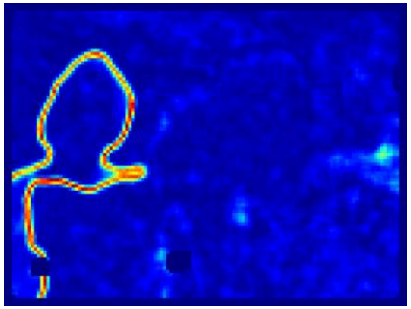
Image

Reference Mask

Noiseprint

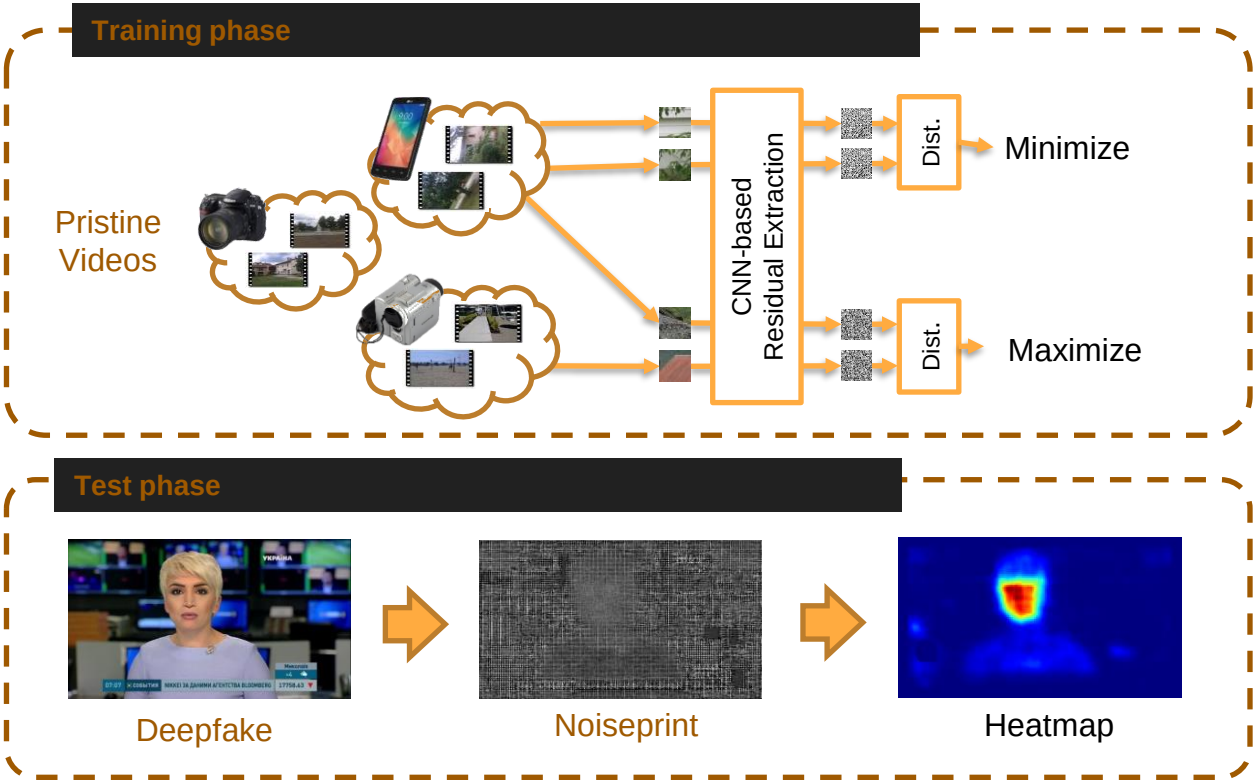






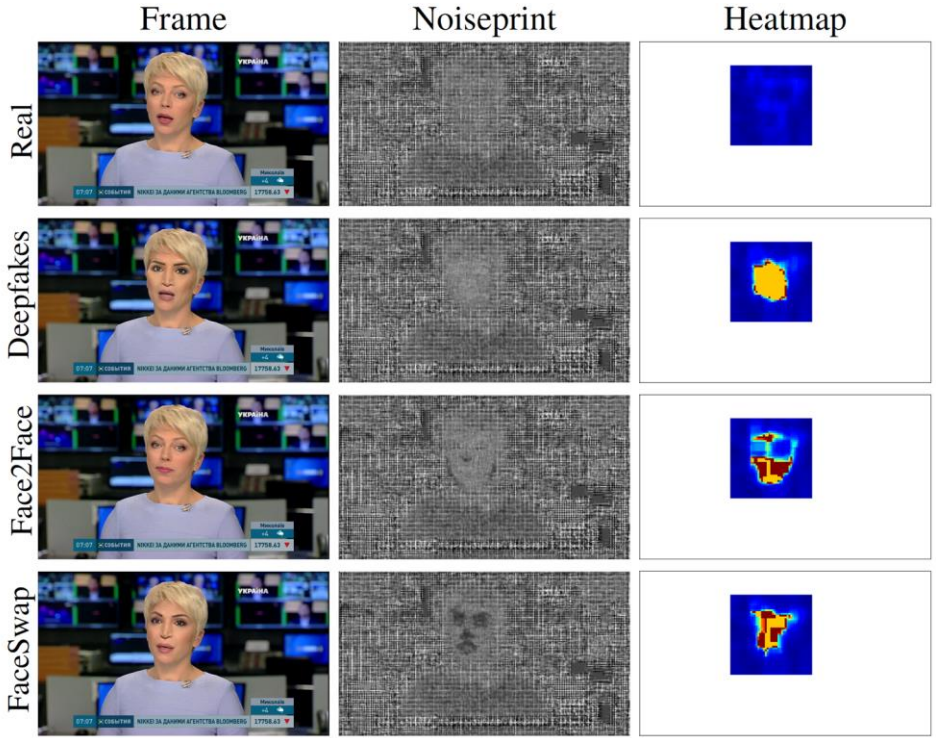
Different ways of clustering

Noiseprint: extension to videos



D. Cozzolino, G. Poggi and L. Verdoliva, “Extracting camera-based fingerprints for video forensics”, CVPR Workshops 2019.

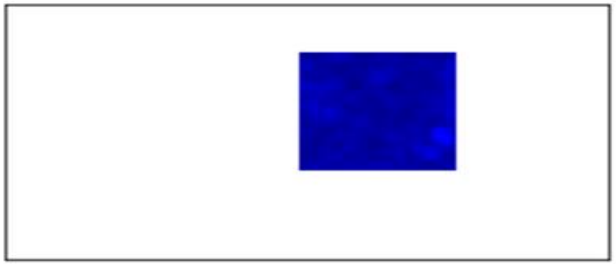
Analysis of different manipulations

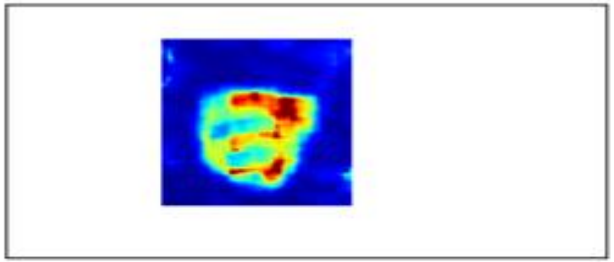


D. Cozzolino, G.Poggi, L. Verdoliva, “Extracting camera-based fingerprints for video forensics,” *CVPR Workshops*, June 2019.

On YouTube

Face analysis



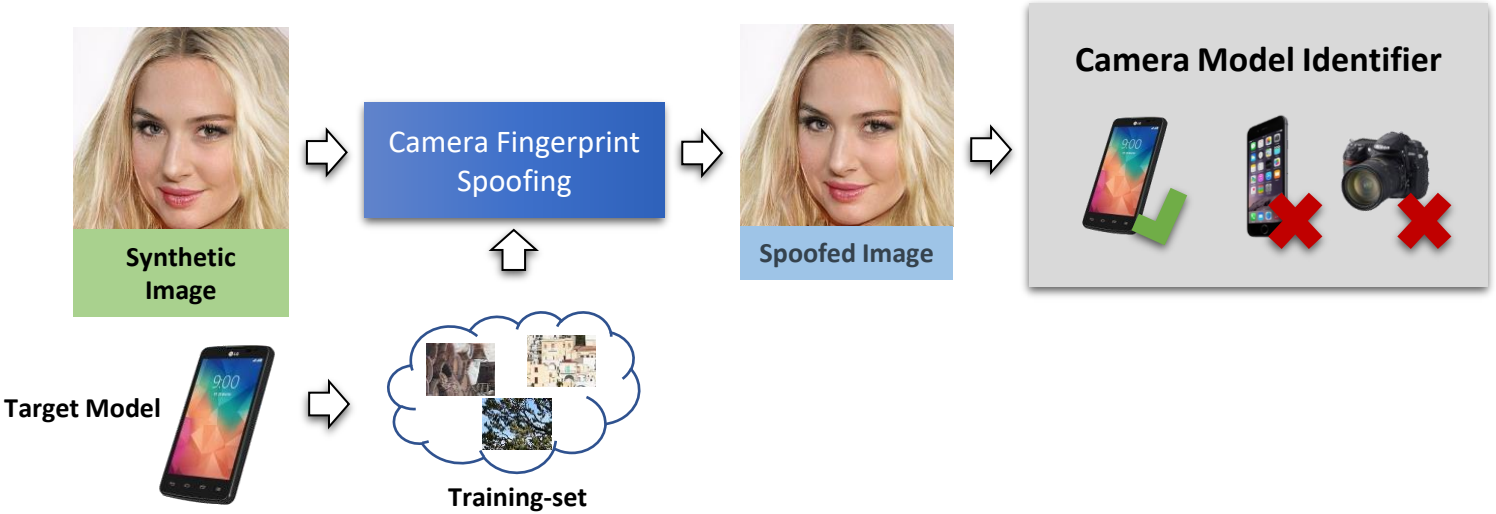


Adversarial scenario

- Adversarial perturbations to DeepFakes [Huang19, Carlini20, Goebel20, Neekhara20, Wang20]
- GAN fingerprints removal [Tolosana19]
- Camera fingerprints insertion [Cozzolino19]
- Camera/Device anonymization [Andrews20, Chen20, Picetti20]

A possible attack to camera fingerprints

- Make a synthetic image appear like acquired by a real camera



D. Cozzolino, J. Thies, A. Roessler, M. Niessner, L. Verdoliva, “SpoC: Spoofing Camera fingerprints”, *arXiv:1911.12069*, Dec. 2019.

Conclusions

- Technology advances very fast and new and more realistic deepfakes are generated
- Developing reliable forensic detectors is a very hard task (JPEG can help to detect manipulations but can also reduce the artifacts)
- Successful solutions
  - should take into account possible non-malicious post-processing (say, compression)
  - should be able to generalize to new/unseen attacks
  - account for skilled attackers who know the principles on which forensics detectors rely

Future directions

- Need to characterize the malicious intent in the detection process
- Need of interpretable solutions
- Multimodal analysis
- Active methods
- Possible integration of JPEG standard and detection algorithms