Vision and Perception 2022-2023

Computer vision forensics and security: media forensics and deepfake detection

ALCOR Lab

SAPIENZA
UNIVERSITÀ DI ROMA

References

- Multimedia forensics book – Springer https://link.springer.com/book/10.1007/978-981-16-7621-5
- Deep Learning for Multimedia Forensics, Foundations and Trends in Computer Graphics and Vision, 12, 4, 309-457, 2021 http://aris.me/pubs/forensics-survey.pdf

Overview

- Deep learning and computer vision (image/video processing) for multimedia forensics
  - Source identification
  - Forgery detection
  - Application scenarios

- Objectives:
  - To provide an insight within the scientific thematic
  - To present some main techniques
  - To introduce the principal application scenarios

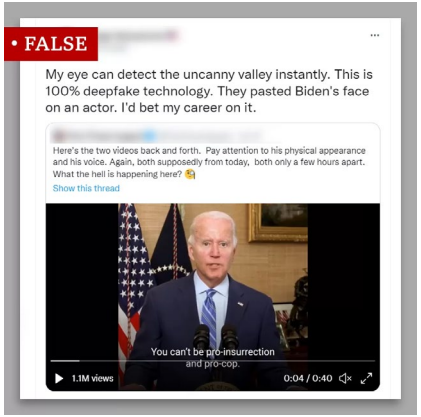
What is Computational/Digital Forensics?

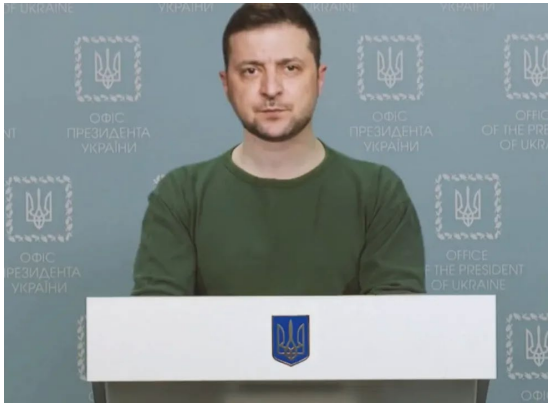
- «Collection of scientific techniques for the **acquisition, preservation, validation, identification, analysis, interpretation, documentation and presentation** of digital evidence derived from digital sources for the purpose to facilitating or furthering the reconstruction of events, usually of a criminal nature.»

Edward Delp- Purdue University

Research context – Why?

- Digital media strongly contribute to the viral diffusion of information through social media and web channels, and play a fundamental role in the digital life of individuals and societies
- Growing issue about the trustworthiness of digital media

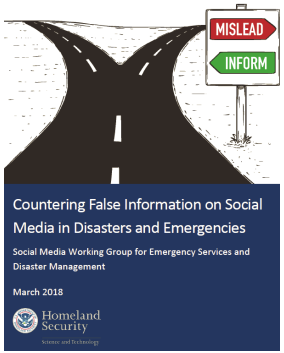




Research context – What?

- Weaponized information and information warfare
- Coordinated propaganda campaigns
- Spreading of disinformation and disinformation detection
- Fake news detection and fact-checking
- Forensics analysis in a court of law

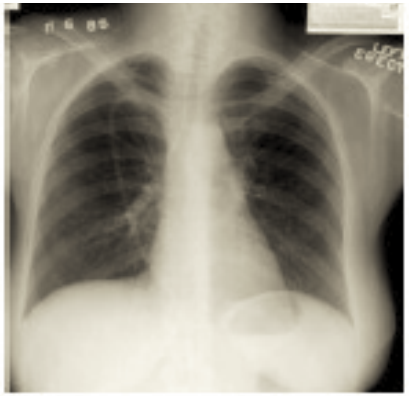
"NewsWire: April 1, 2019, Bob Smith On a rainy spring day, a vast, **violent group** gathered in front of the US Capitol to **protest** recent cuts in Social Security."



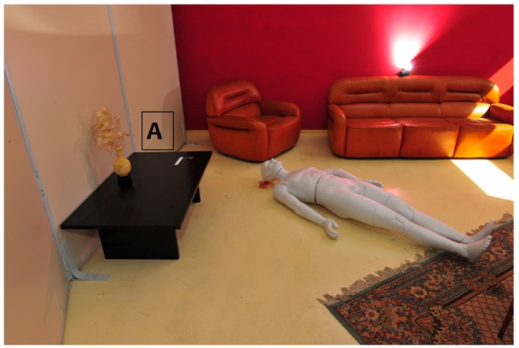
In a trial

Reputation attacks

Insurance frauds



Crime scene alterations

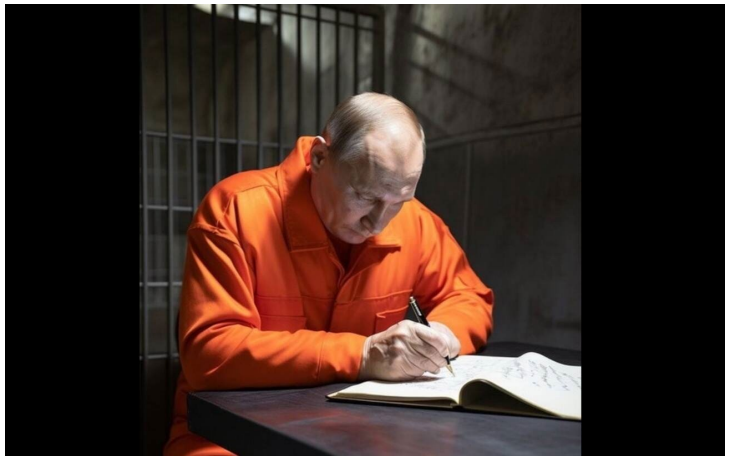


A new threat

- Deep Fakes
- New media-synthesis methods

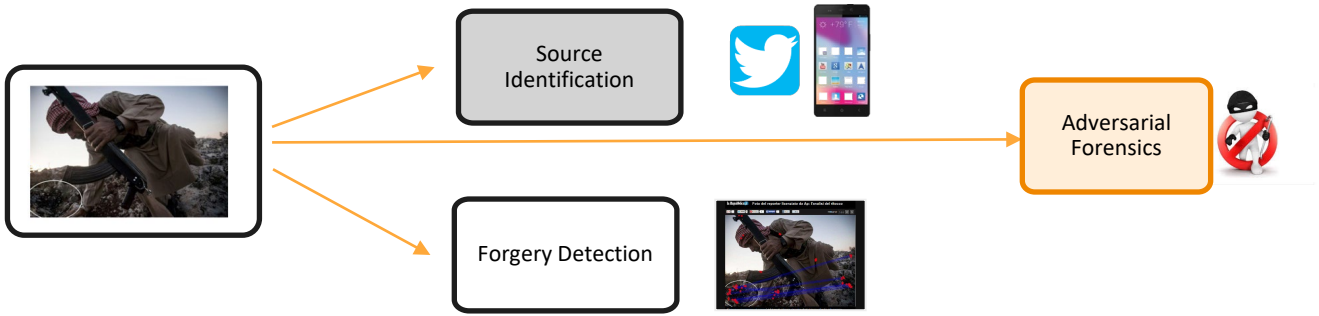
reddit

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



Multimedia Forensics

- **Origin**
  - Source identification, link multimedia content to a particular device or social network
- **Authenticity**
  - Forgery detection, deciding on the integrity of the media (image, video or audio)
- **Security**
  - Adversarial forensics/Counter forensics

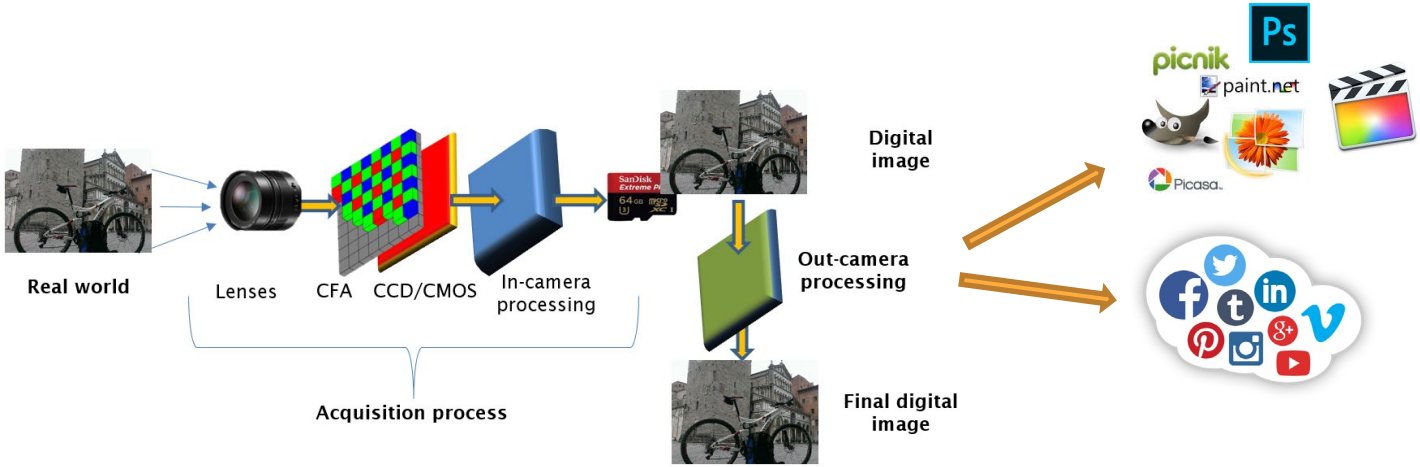


Amerini I., Anagnostopoulos A., Maiano L., Ricciardi Celsi L., *Deep Learning for Multimedia Forensics*, Foundations and Trends in Computer Graphics and Vision, 12, 4, 309-457, 2021
I. Amerini, G. Baldini and F. Leotta (Eds.), *"Image and Video Forensics"*, ISBN 978-3-0365-2806-9 (Hbk); ISBN 978-3-0365-2807-6 https://doi.org/10.3390/books978-3-0365-2807-6, January 2022

Multimedia forensics: the rationale

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

Each phase leaves distinctive footprints!
- at the signal level
- at the metadata/file container level



Basic principles

- Only the image (video) and sometimes the device in our hands.
- No external information like metadata.

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

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

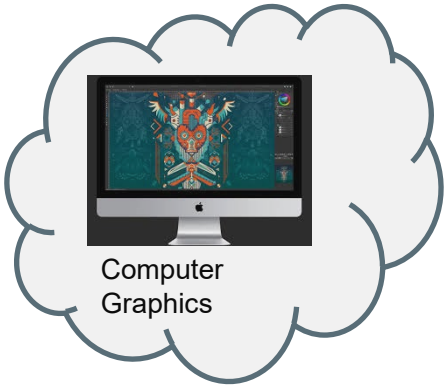
  - *Fingerprint extraction*
  - *Fingerprint classification*

PART 1

Source Identification

Source identification

- Which **CLASS** of devices



Source identification

- Which **BRAND/MODEL**

Nikon?
Canon?
Sony?


Nikon D70
Nikon D3300


Canon eos 1300d
Canon ixus 115 HS

Sony cyber-shot dsc-h300
Sony a6000

Source identification

- Which **DEVICE**

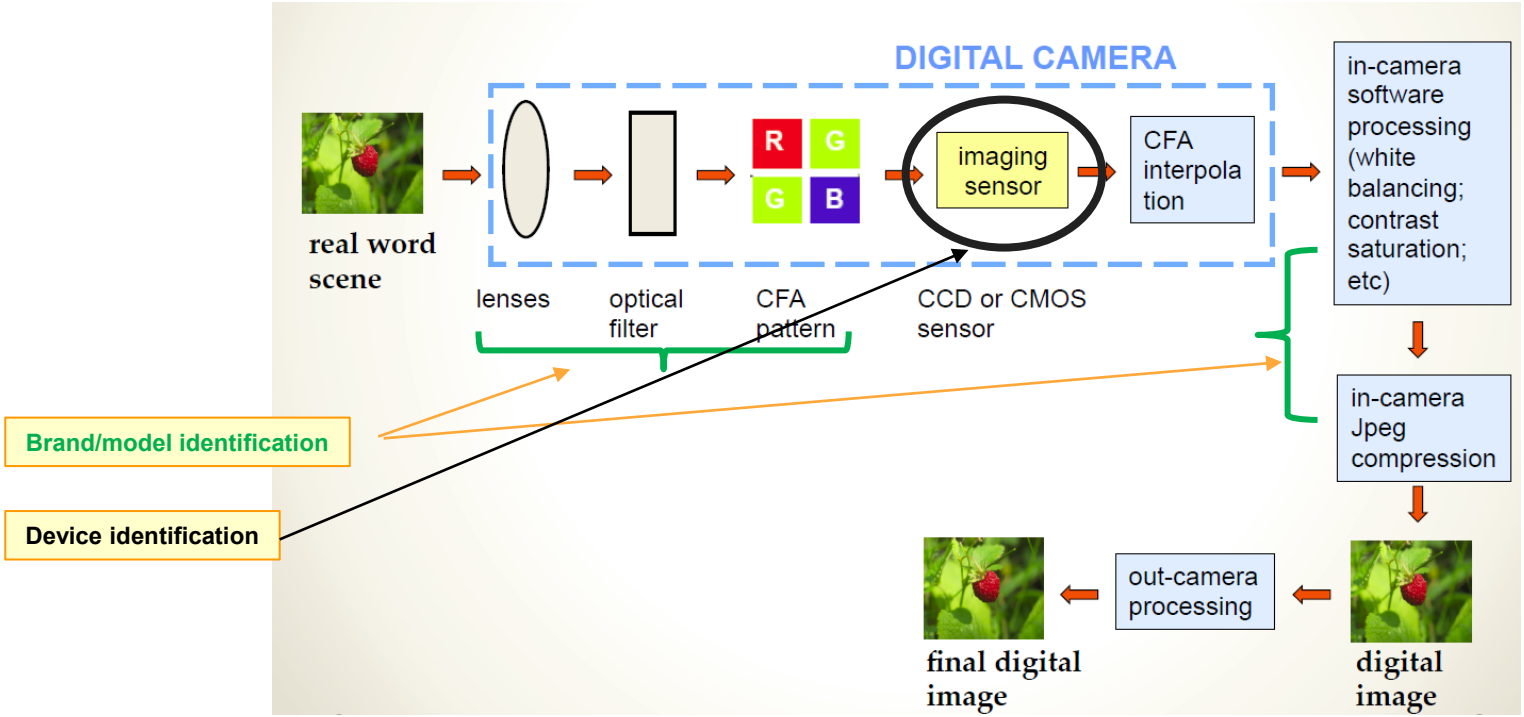
Which Nikon D3300?

Serial Number
000111201

Serial Number
000111204

Serial Number
000111207

The acquisition process (in detail)

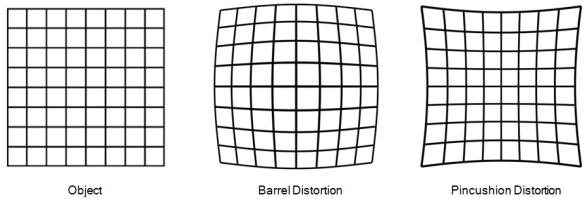


Lens distortion

- Radial distortion:
  since the lens is non-ideal, lines becomes curves in the image plane

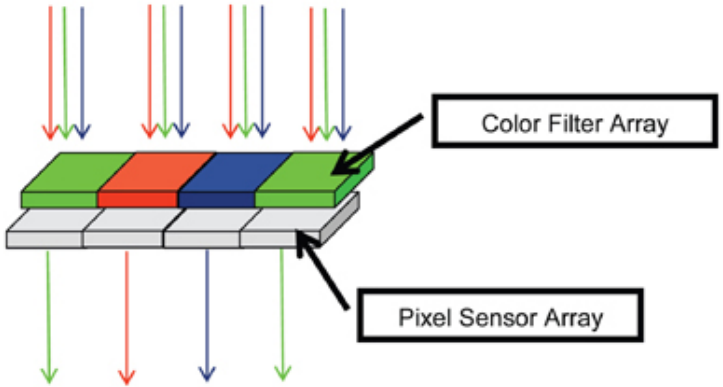
  - Distortion properties and parameters can be estimated and used as a signature

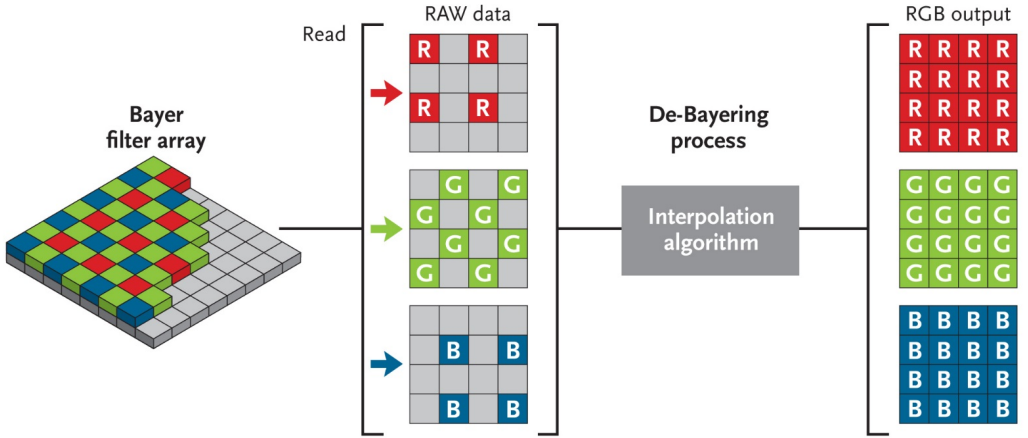
  - Main problem: lenses can be changed on DSLR cameras



Color Filter Array

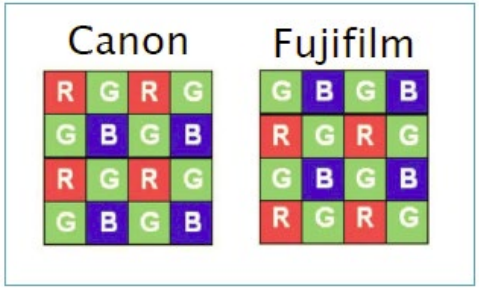
- CFA is a thin film that selectively permits a certain component of light to pass towards the sensor.
- In practice, for each pixel only one particular primary color is gathered. The sensor output is successively interpolated to obtain all the 3 colors for each pixel (demosaicking)





Color Filter Array

- Different producers may use:
  - Different CFA patterns
  - Proprietary demosaicking algorithms
    - Bilinear
    - Bi-cubic
    - Median filter
    - Gradient based
    - Adaptive color plane

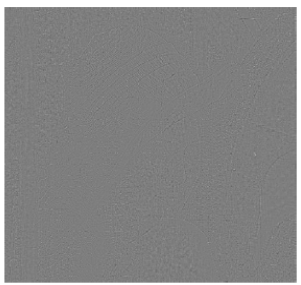


- Methods exist to blindly estimate the CFA pattern and learn neighboring pixel relationships

CCD sensor imperfections

- PRNU (Photo Response Non Uniformity Noise) is caused by **the different *sensitivity* of the sensors to light**
  - Due to the manufacturing process
  - Does not depend on temperature and time

- **If we capture this noise pattern, we can create a distinctive link between a camera and its photos**



[Fridrich at Al, TIFS 2006]

PRNU fingerprint model

A digital image **I** taken from camera C can be modeled as

$$I' = I + IK + \theta$$

Acquired image

Denoised image

Other noise terms (shot, readout etc..)

PRNU fingerprint

$$\hat{K} = \frac{\sum_{i=1}^{N} W_i I'_i}{\sum_{i=1}^{N} (I'_i)^2} \qquad W_i = I' - I_F \qquad I_F = F(I') \approx I$$

Observation: The PRNU pattern noise is a **multiplicative noise**

PRNU fingerprint detection

- Let **Y** be an input image (from the same camera C or another one)
- The presence of $\boldsymbol{K}$ in **Y** can be determined by means of the **correlation detector**

$$\boldsymbol{\rho = corr(W_Y, \widehat{K}Y)}$$

Noise residual of image *Y*

Reference fingeprint

where:

$$corr(X,Y) = \frac{(X - \overline{X}) \cdot (Y - \overline{Y})}{\|X - \overline{X}\| \, \|Y - \overline{Y}\|}$$

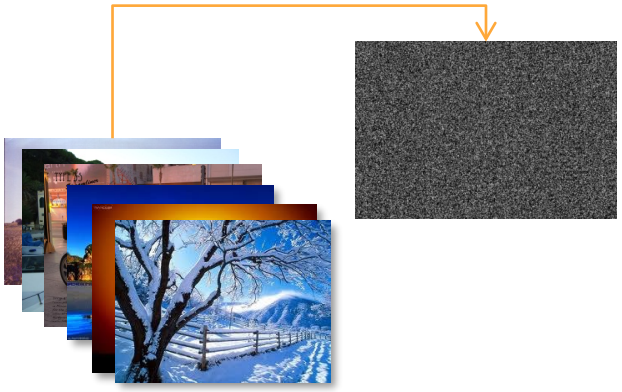
$$X \cdot Y = \sum_{i,j} X[i,j]Y[i,j]$$

$$||X|| = \sqrt{X \cdot X}$$

High when **Y** was acquired by camera C with PRNU fingerprint **K**, low otherwise

PRNU estimation

- For getting a good estimate of the PRNU fingerprint
  - flat-field images (uniform content, low variance)
  - The luminance as high as possible but not saturated
  - About 20 images are sufficient
- The estimated PRNU fingerprint for camera C can be used for testing purposes.

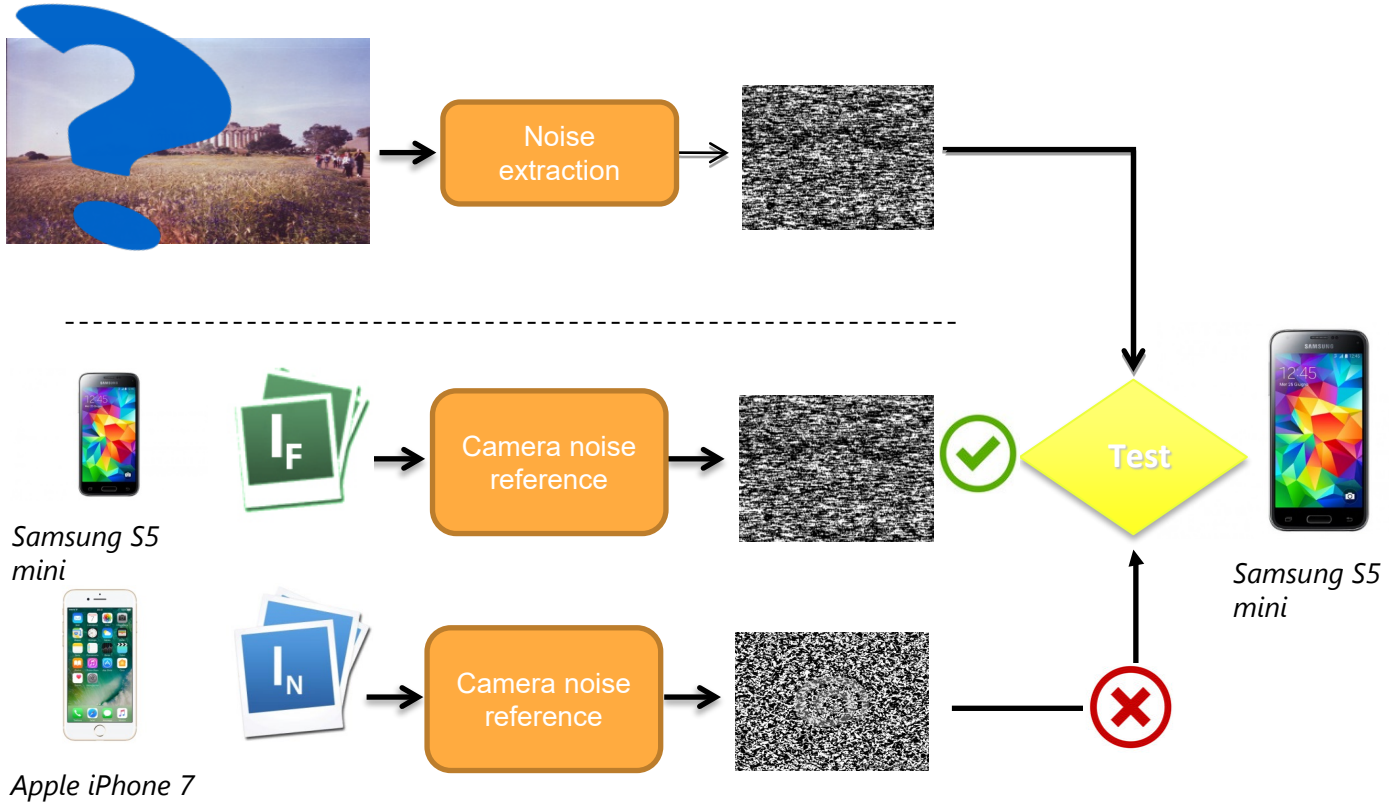


PRNU correlation test

- Which camera does the test image come from?

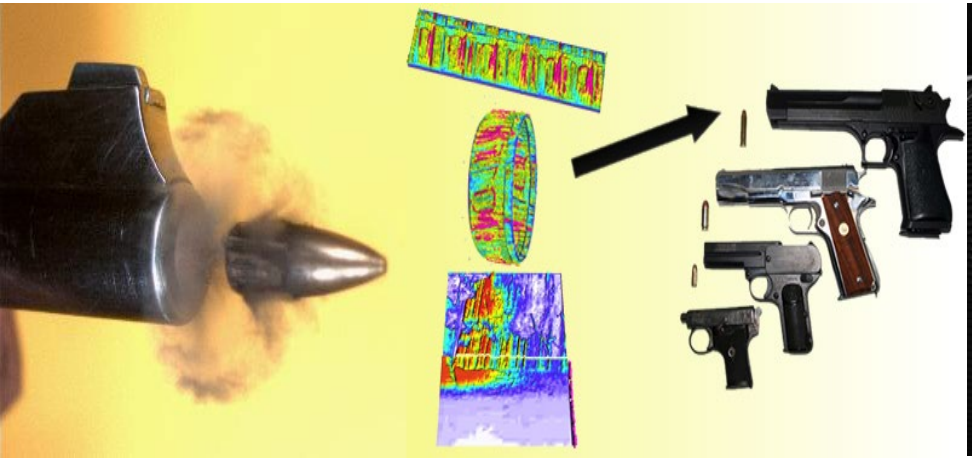
- The PRNU of the tested image is extracted and correlated with each camera's reference PRNU pattern.
- The camera with the highest correlation is the one that acquired the image

PRNU correlation test



A well known analogy

Firearms Identification



Digital Cameras Identification

Social Network identification

- In general, **source identification** is the process to link a multimedia content to a particular **acquisition device.**

- **Lately source identification also refers to establish the social network of origin (platform provenance).**

Analysis of the social media of origin

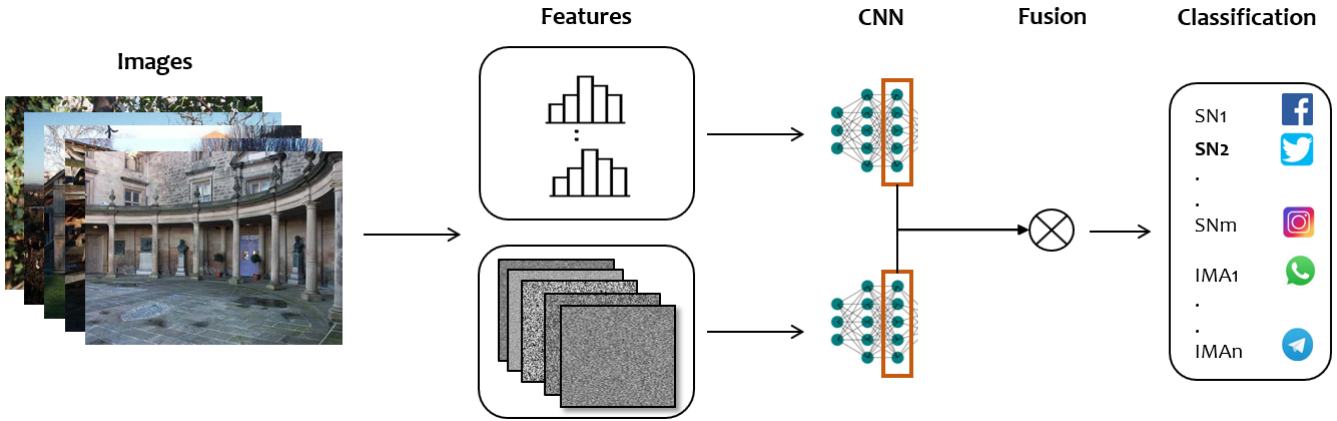
- When we upload a content to a social-media platform, it usually goes through a series of operations, which most commonly may include recompression and a resize
- Current state of the art focuses on the **analysis of images** and commonly rely on:
    - **Discrete Cosine Transform (DCT):** histograms of DCT coefficients allow to distinguish social networks in base of recompression
    - **Noise residual** based on Photo Response Non-Uniformity ($I' = I + IK + \theta$): caused by the different sensitivity of the sensors to light
- A SN will apply a unique modification on every processed image.
- If such consistency and uniqueness are observed, we will know that the traces due to the modification can be exploited for social network identification

Amerini et al., *Social network identification through image classification with CNN*, IEEE Access, vol. 7, pp. 35264–35273, 2019.
Phan et al., *Tracking multiple image sharing on social networks*, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019.
Caldelli et al., *PRNU-based image classification of origin social network with CNN*, in 2018 26th European Signal Processing Conference (EUSIPCO)
Hosler and Stamm, *Primary and secondary social media source identification*, 2021. arXiv: 2105.02306.

Social Network Provenance: on image content

FusionNET: CNN-based framework for addressing the social network and instant messaging app identification

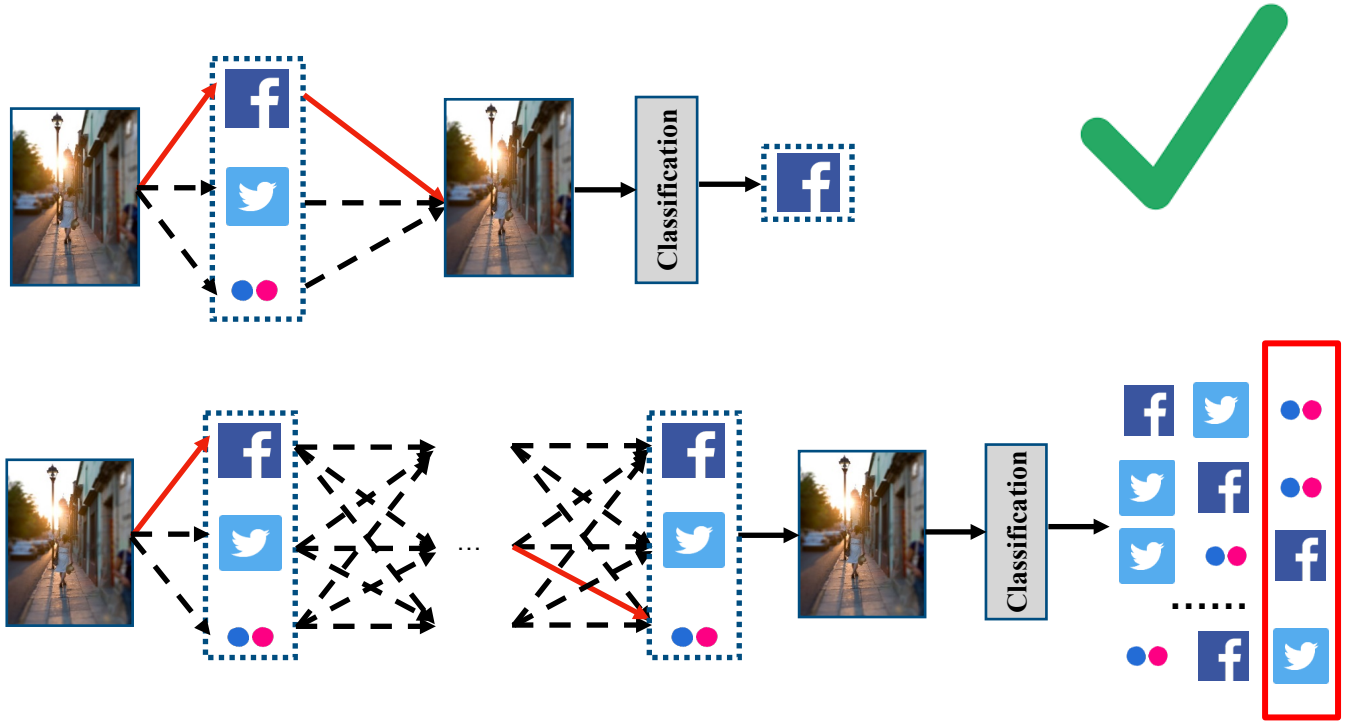
- Dual-modal features for image representation: the histogram of DCT and the sensor noise residuals
- Two CNN branches fed with the respective feature modalities to pull out activation vectors
- Fusion of activation vectors
- Classification of source SNs and IMAs of the images in question.



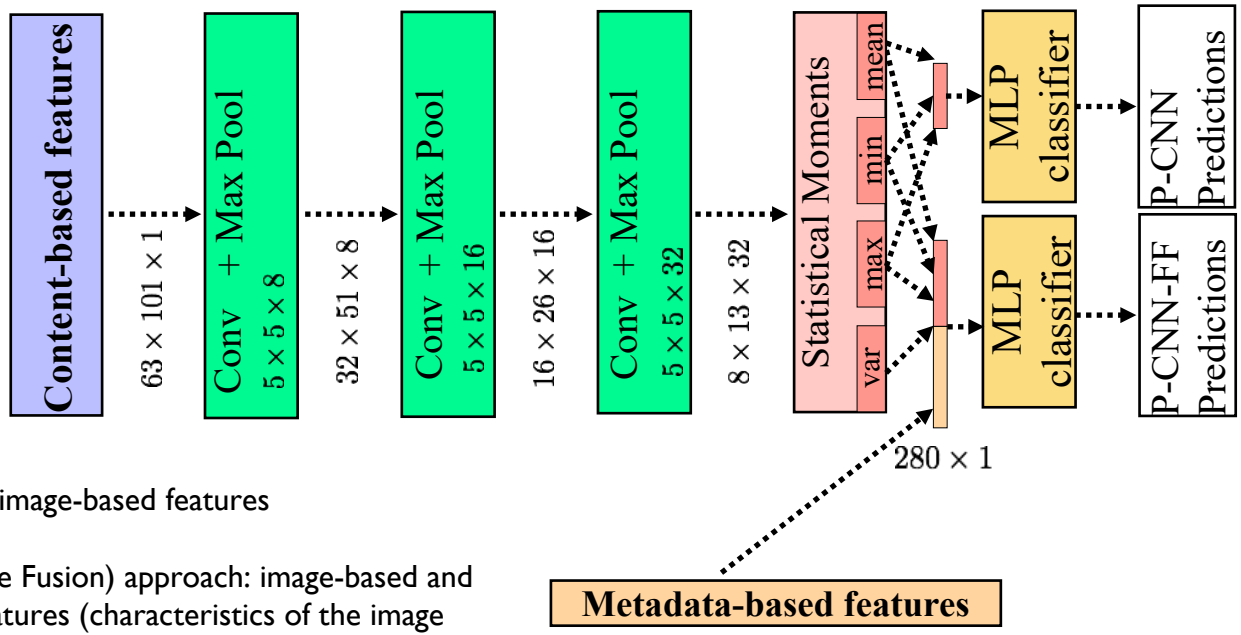
I. Amerini et Al, "Social Network Identification through Image Classification with CNN", IEEE Access 2019
I. Amerini et Al, "Image origin classification based on social network provenance", IEEE TIFS 2017

ENDEAVOUR Scholarships and Fellowships
Australian Government
Charles Sturt University

Single/Multiple shares



Multiple up-down classification

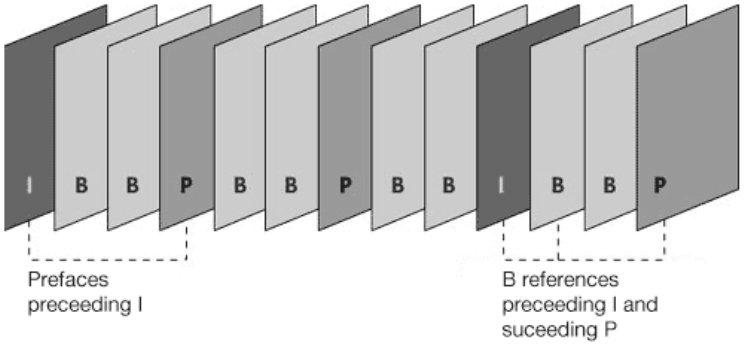


- P-CNN approach: image-based features
- P-CNN-FF (Feature Fusion) approach: image-based and metadata-based features (characteristics of the image file e.g. quantization tables, image size).

Q.-T. Phan, G. Boato, R. Caldelli, I. Amerini , "Tracking Multiple Image Sharing On Social Networks", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.

Social Network Provenance: on video content

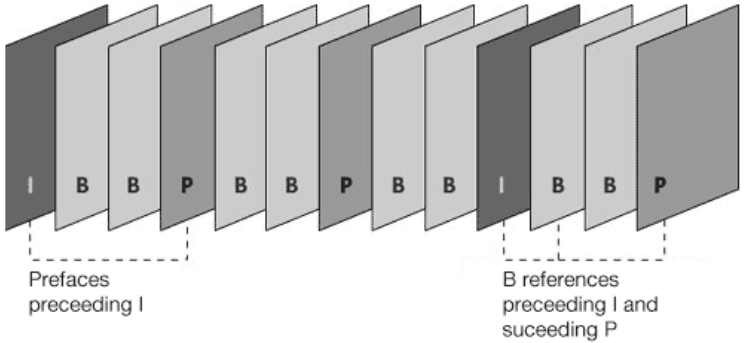
The upload may include recompression, a resize, and in some cases the removal of some frames to reduce the band-width requirement.



H.264 video sequences

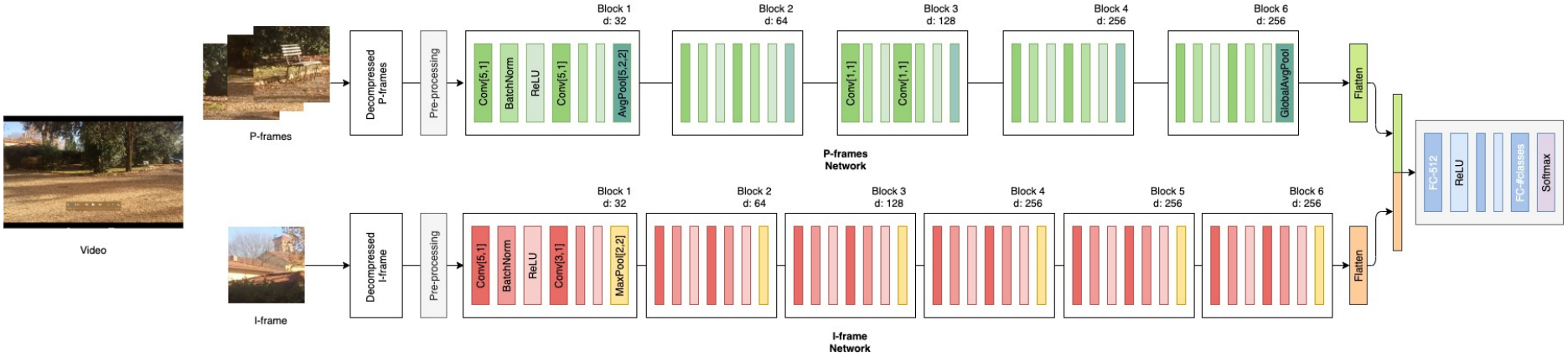
In video coding, a video is represented as a sequence of *groups of pictures* (GOP)s, each of which is made of:

- **I-frames**, not predicted from any other frame and are independently encoded using a process similar to JPEG compression.
- **P-frames**, predictively encoded using from the previous I-frame through motion estimation and compensation
- **B-frames**, similar to P-frames but obtained from both neighboring I- or P- frames for motion estimation



Detecting videos shared through social media

- Convert frames from RGB to YUV
- Preprocess the Y channel of I-frames and P-frames by means of high pass filtering
- Features are extracted by an I-frame network and P-frame network streams

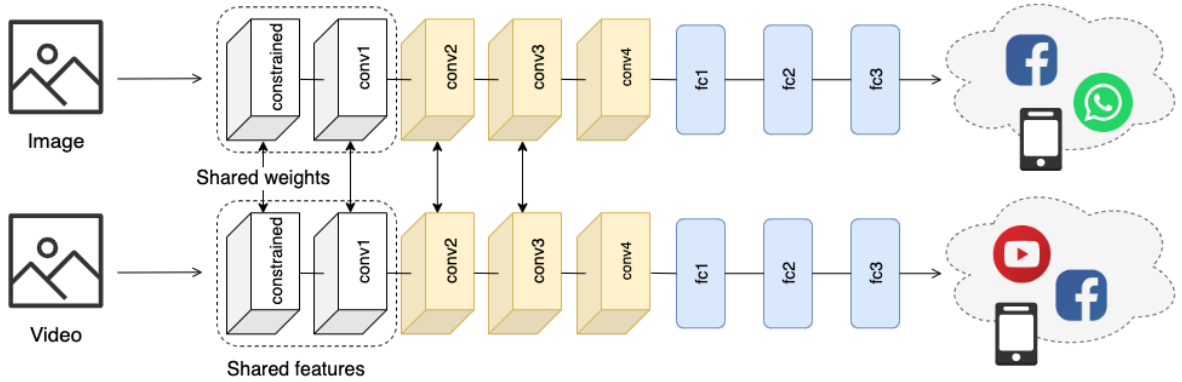


Amerini I., Anagnostopoulos A., Maiano L., Ricciardi Celsi L., *Learning double-compression video fingerprints left from social-media platforms*, in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

Learning from multiple domains

Learning from multiple domains (images and videos):

- a neural network with transfer learning by transferring features from the image domain to the video one
- a **multitask learning model** trained on both images and videos to recognize the social media of origin



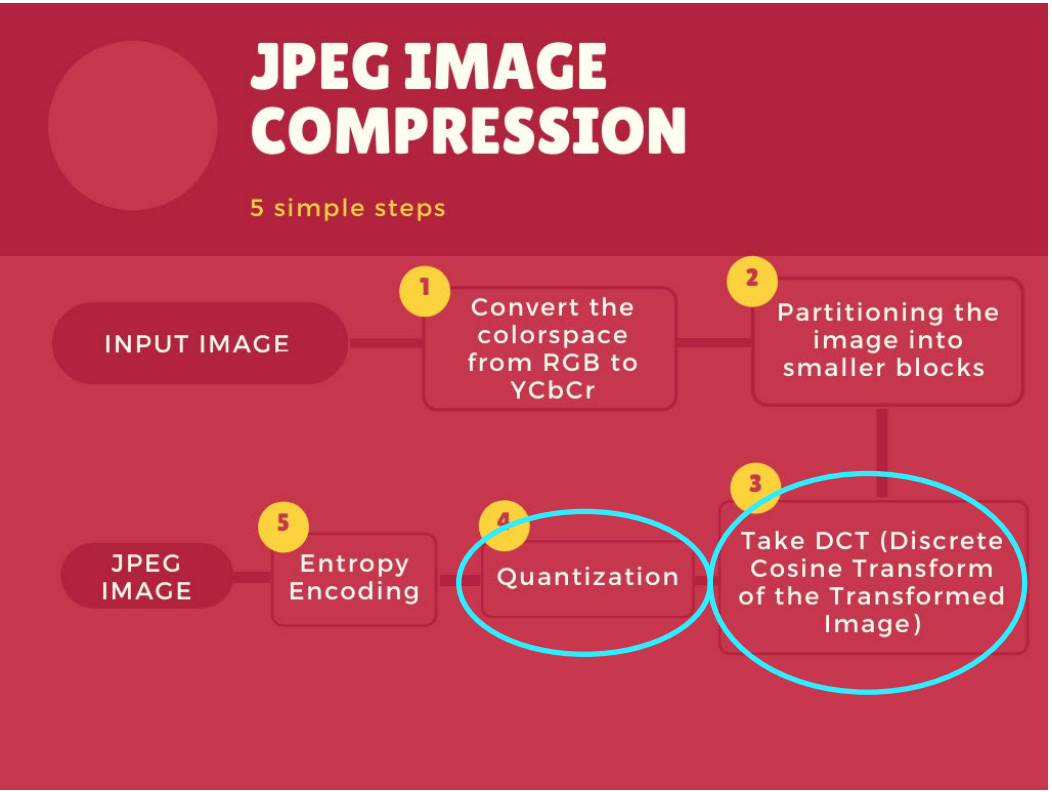
Maiano L., Amerini I., Anagnostopoulos A., Ricciardi Celsi L., *Identification of social media platform of videos through the use of shared features*, Journal of Imaging, vol. 7, no. 8, 2021

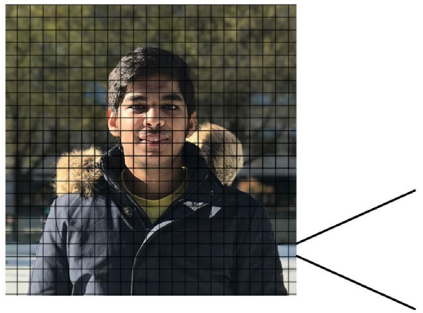
Interlude: JPEG compression and DCT

JPEG compression footprints

- Like any other image processing, JPEG leaves traces into the image, especially at low quality factors
- Such traces can be exploited to gather useful information on the image
- Some JPEG artifacts are immediately identified
    - Blocking due to block discontinuities
    - Ringing on edges due to the DCT
    - Graininess due to coarse quantization
    - Blurring due to high frequency removal
- Other (statistical) alterations are way more subtle to identify!

JPEG baseline encoder



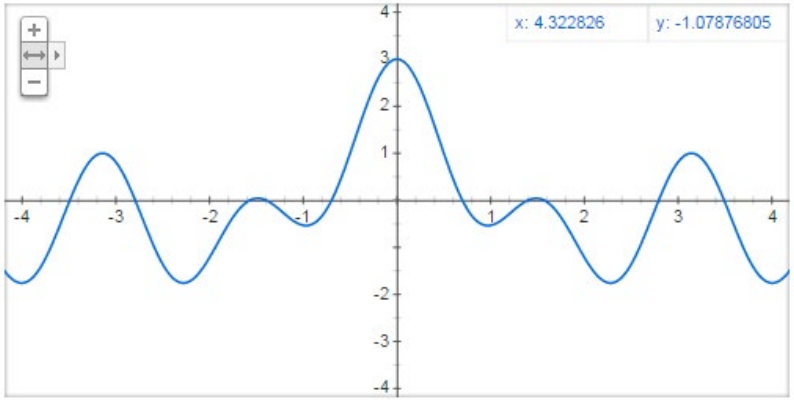


Discrete Cosine Transformation Matrix

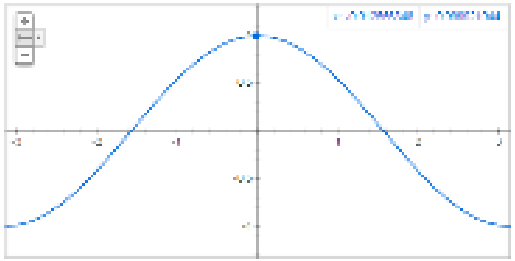
$$\begin{bmatrix} -415 & -33 & -58 & 35 & 58 & -51 & -15 & -12 \\ 5 & -34 & 49 & 18 & 27 & 1 & -5 & 3 \\ -46 & 14 & 80 & -35 & -50 & 19 & 7 & -18 \\ -53 & 21 & 34 & -20 & 2 & 34 & 36 & 12 \\ 9 & -2 & 9 & -5 & -32 & -15 & 45 & 37 \\ -8 & 15 & -16 & 7 & -8 & 11 & 4 & 7 \\ 19 & -28 & -2 & -26 & -2 & 7 & -44 & -21 \\ 18 & 25 & -12 & -44 & 35 & 48 & -37 & -3 \end{bmatrix}$$

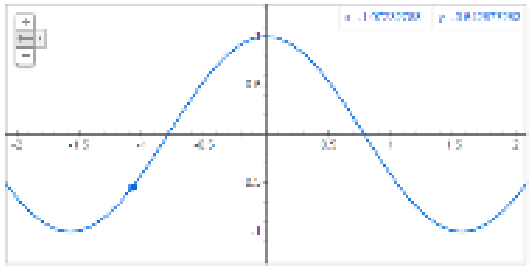
Discrete cosine transform

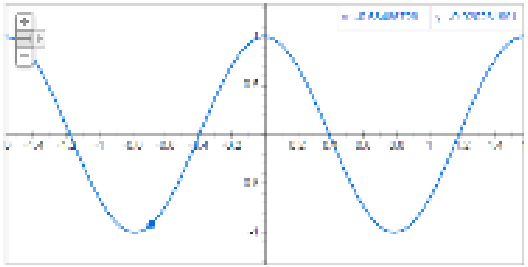
- Any numeric signal can be recreated using a combination of cosine functions.



sum of cos(x)+cos(2x)+cos(4x)

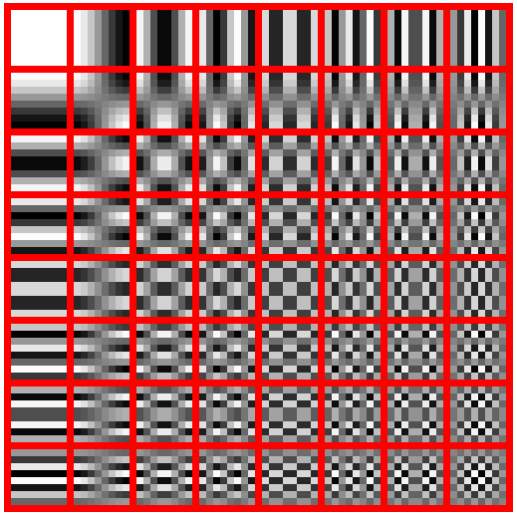






Discrete cosine transform

- Images are split into 8x8 blocks and reconstructed using cosine basis functions.

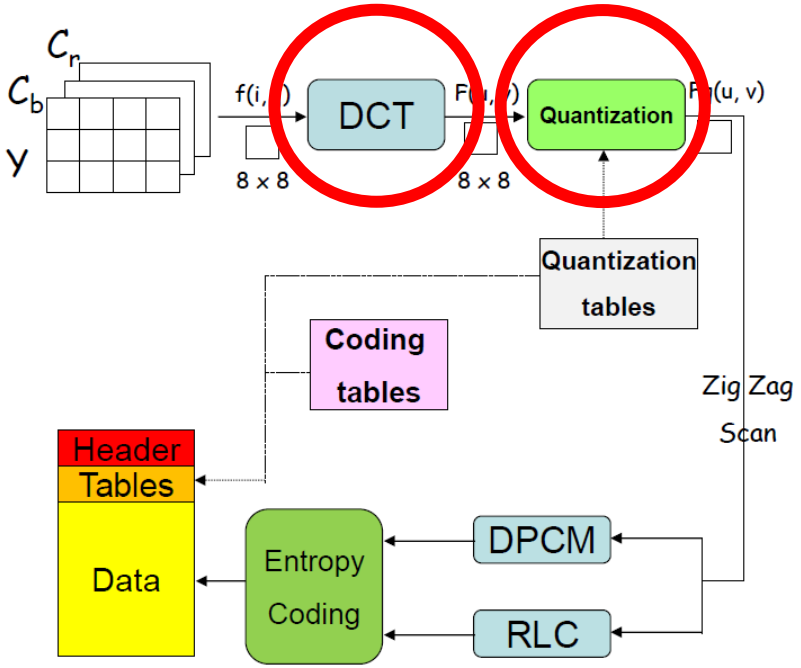


From spatial domain to frequency domain

- Original Lena image

- 2D DCT

JPEG baseline encoding

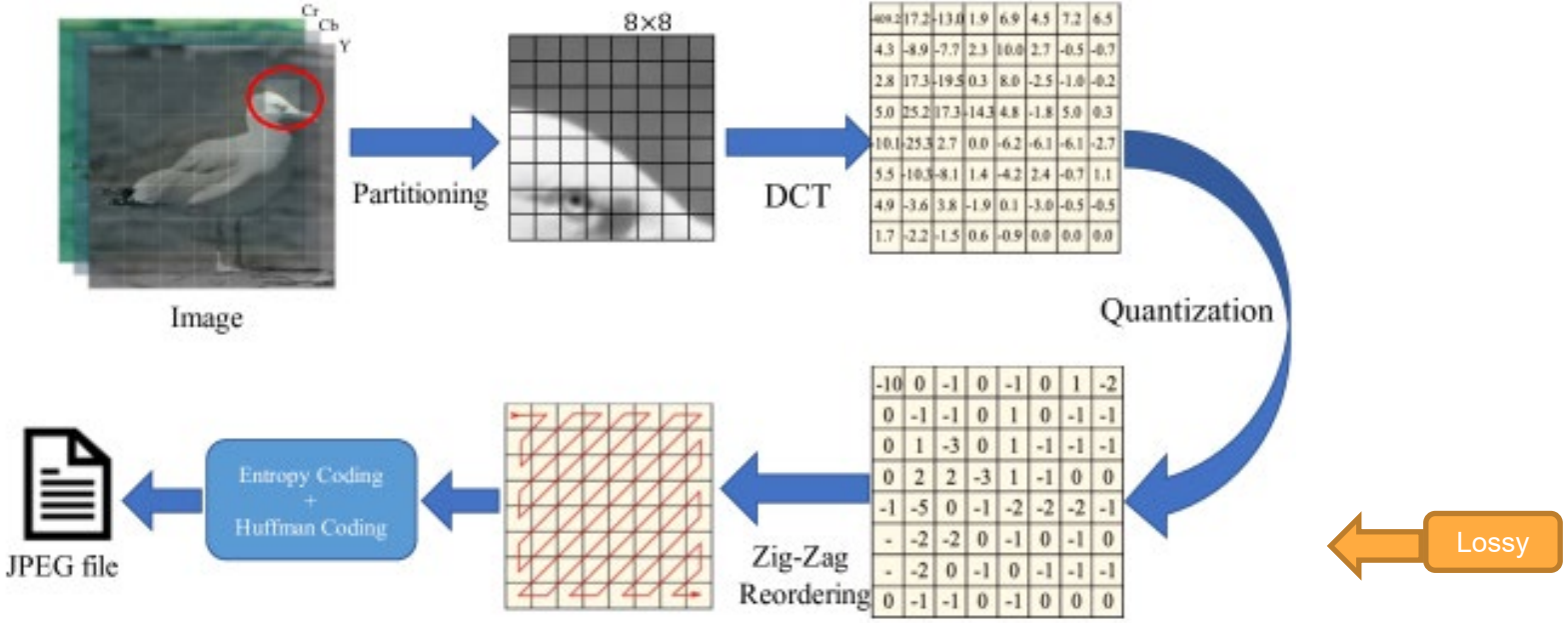


Main steps:

1. Discrete Cosine Transform of each 8x8 pixel block
2. Scalar quantization — Lossy
3. Zig-zag scan to exploit redundancy
4. Differential Pulse Code Modulation (DPCM) on the DC component and Run Length Encoding of the AC components
5. Entropy coding (Huffman)

Reverse order for decoding

JPEG baseline encoding



Quantization phase

- Quantize the 8x8 DCT coeff matrices using a quantizing matrix Q for each channel

$$F_q(u,v) = \left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor, \quad \hat{F}(u,v) = F_q(u,v) \cdot Q(u,v)$$

$$Err(u,v) = \hat{F}(u,v) - F_q(u,v)$$

Q(u,v), *quantization step* at frequency (u,v)

a. Low compression

1	1	1	1	1	2	2	4
1	1	1	1	1	2	2	4
1	1	1	1	2	2	2	4
1	1	1	1	2	2	4	8
1	1	2	2	2	2	4	8
2	2	2	2	2	4	8	8
2	2	2	4	4	8	8	16
4	4	4	4	8	8	16	16

b. High compression

1	2	4	8	16	32	64	128
2	4	4	8	16	32	64	128
4	4	8	16	32	64	128	128
8	8	16	32	64	128	128	256
16	16	32	64	128	128	256	256
32	32	64	128	128	256	256	256
64	64	128	128	256	256	256	256
128	128	128	256	256	256	256	256

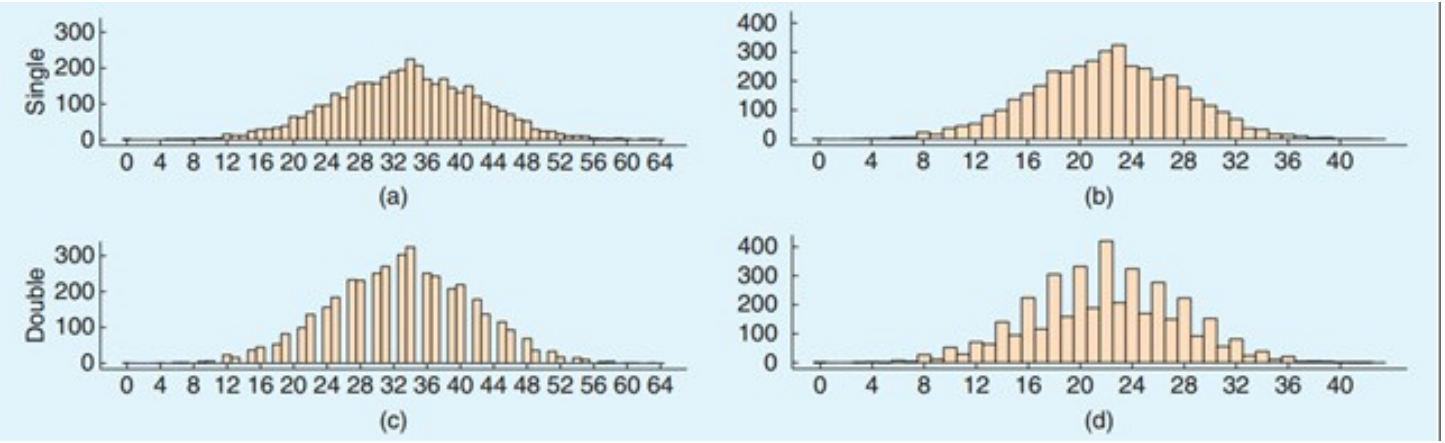
- Eye is more sensitive to low frequencies (upper left corner of the 8x8 DCT block) less sensitive to high frequencies (lower right corner)
    - Idea: quantize more (large quantization step) the high frequencies, less the low frequencies
- The values of Q are controlled with a parameter called Quality Factor (QF) which ranges from 100 (best quality) to 1 (extremely low)

Double JPG compression

- Double JPEG compression is when an image is JPEG compressed first with QF1 and then JPEG compressed again with QF2

- Statistical footprints, due to double quantization
    - Several approaches have been proposed to reveal the footprints (periodic artifacts) left by double compression

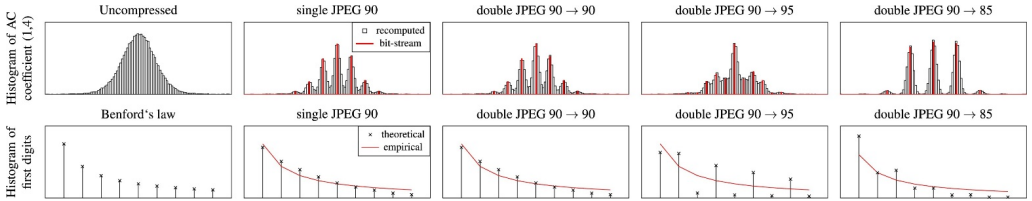
Double JPEG compression



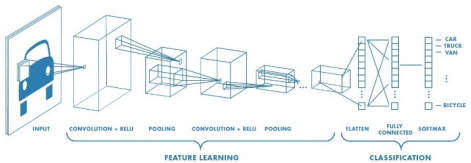
Double compression introduces periodic artifacts in the quantized DCT coefficient histogram

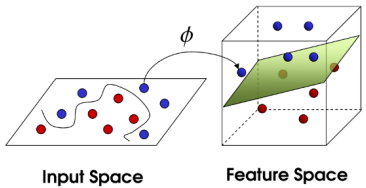
Double JPEG detectors

- Statistical methods based on the histograms analysis



- Use machine learning techniques to build a detector that can distinguish between single quantized histograms ("without holes") and double quantized histograms (with "holes")
  - SVM
  - CNN





Double JPEG compression in forensics

Why understanding whether an image has been JPEG compressed (quantized) twice is important?

Social network identification

Forgery detection

PART 2

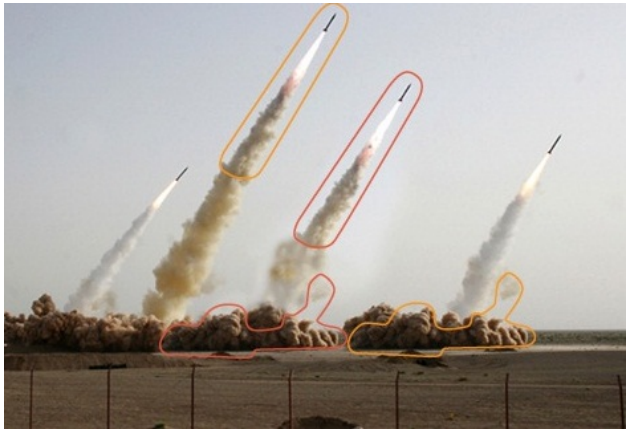
Authenticity verification

Kinds of manipulations

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

Kinds of manipulations

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

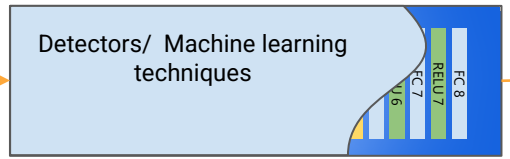


Kinds of manipulations

- Image manipulation categories:
  - Image splicing
  - Copy-Move manipulation
  - Deepfakes-→ (next week)

Forgery detection

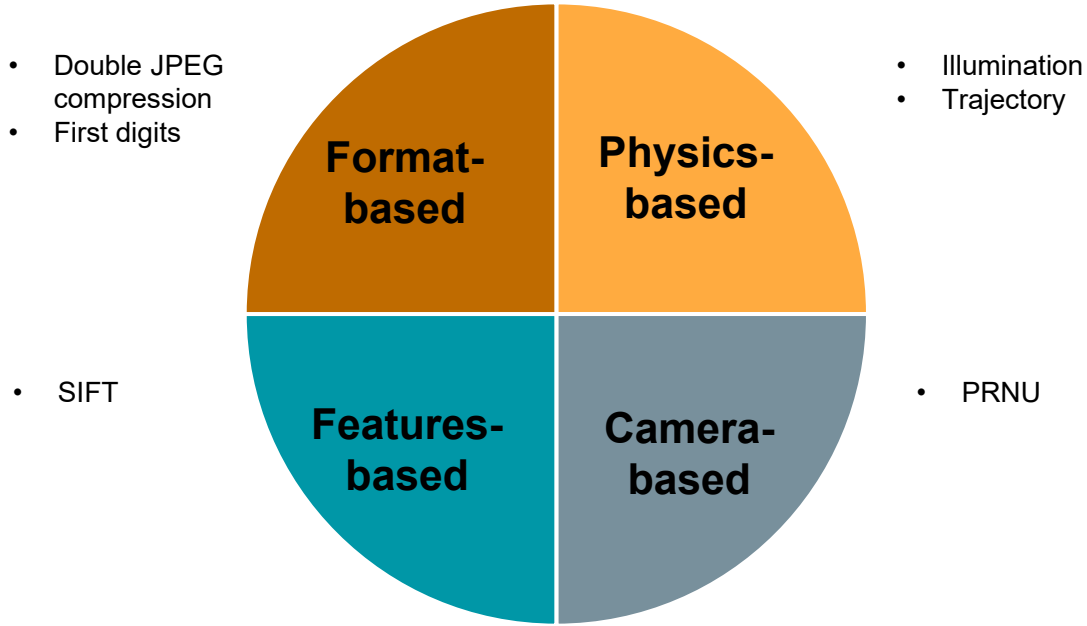
- **Research question:** how a doctored image/video be revealed and localized?
- Given a single probe image, detect if the probe was manipulated and provide mask(s)



The image is doctored with a certain confidence

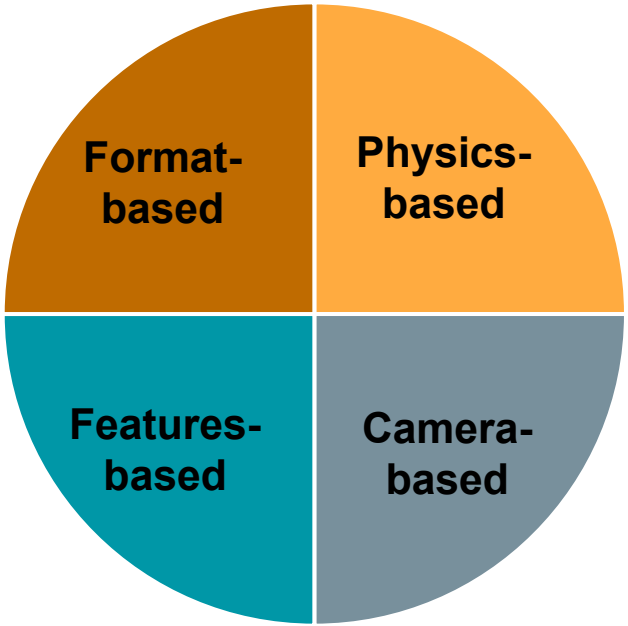
Techniques for tampering detection

- 4 categories

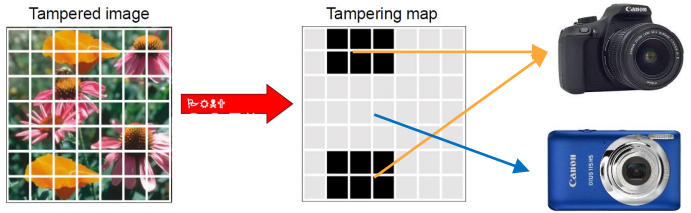


Techniques for tampering detection

- Techniques can roughly be separated in 4 categories

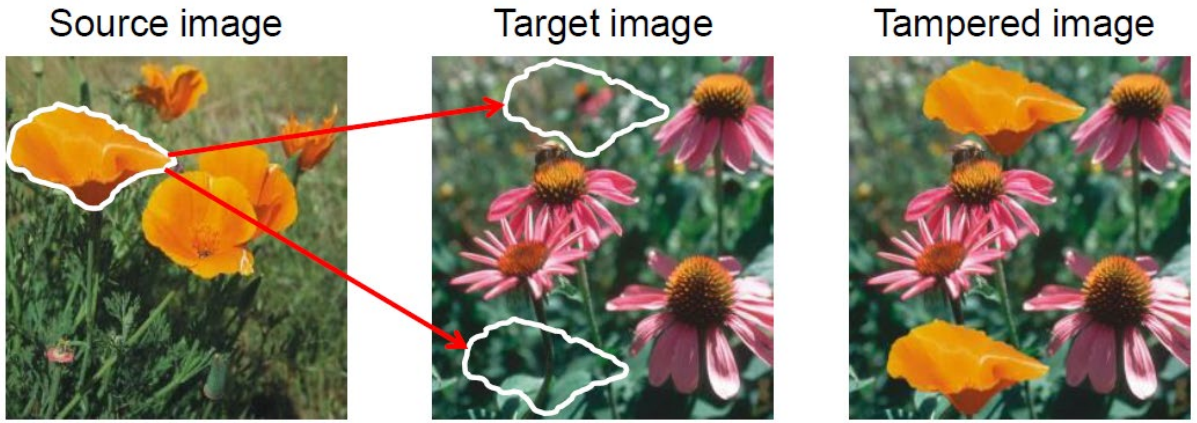


- PRNU



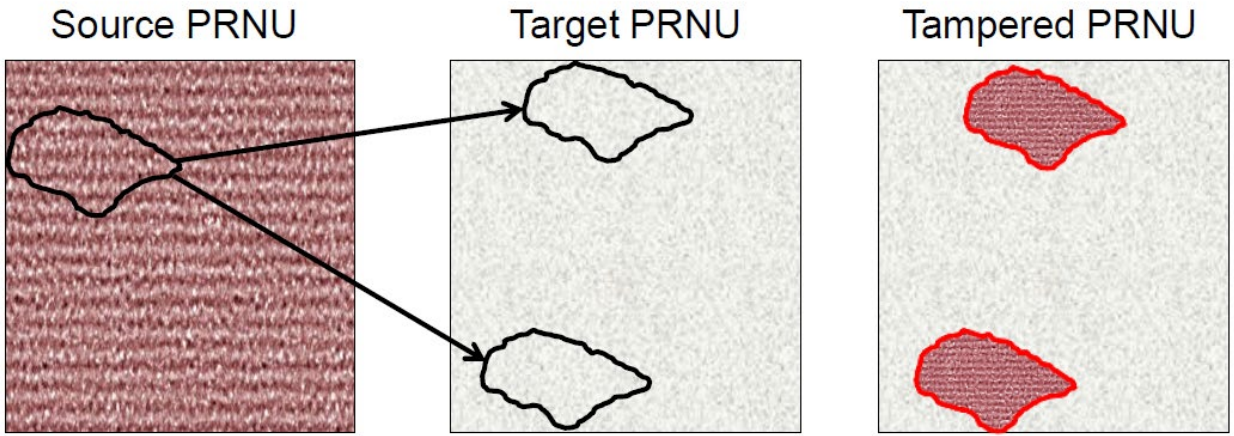
PRNU inconsistencies

- A portion of a different image is selected and copied on the to-be-modified image
- Used to remove or add content to an image



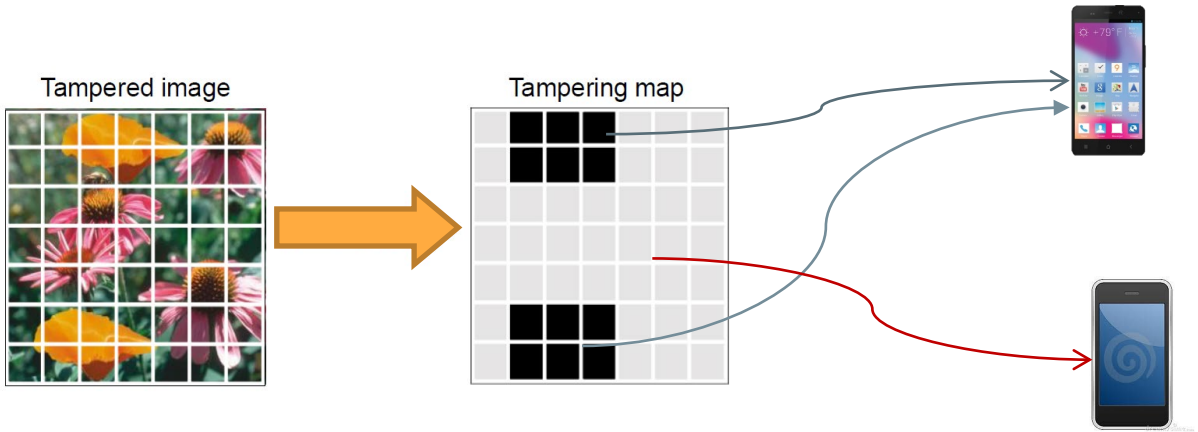
PRNU inconsistencies

- **The source PRNU “behind” the visual content is also copied and pasted on the tampered image**
  - By looking for local *statistical inconsistencies* of the PRNU of the tampered image, it is possible to reveal the manipulation



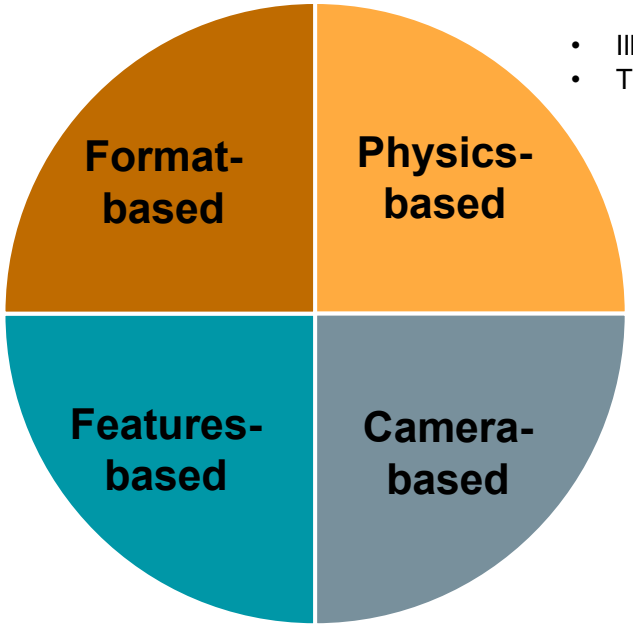
Tampering detection

- The image is divided into blocks (overlapping or not) that are analysed separately
- Each block is labelled as belonging to a **camera C1** or otherwise to an another **camera C2**, leading to a tampering map

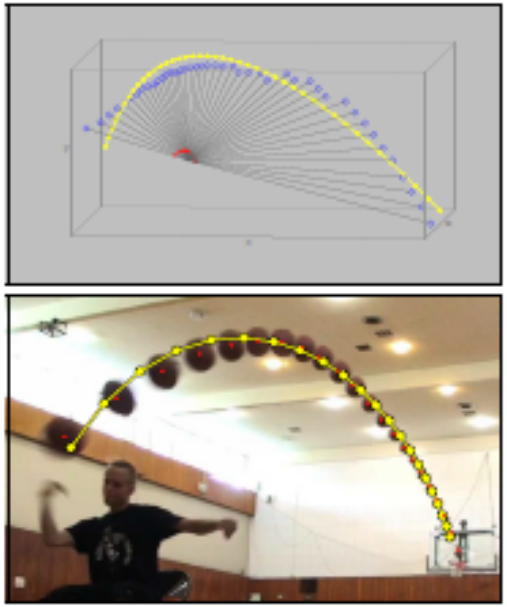


Techniques for tampering detection

- 4 categories

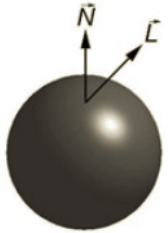


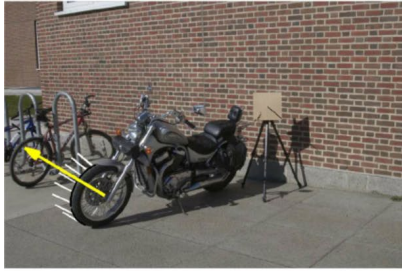
- Illumination
- Trajectory

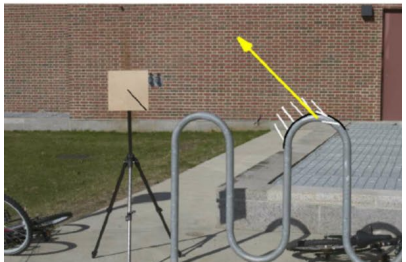


Light inconsistencies

- Under certain assumptions it is possible to derive the direction of light from shadows
    - The surface of interest is Lambertian
    - The surface has a constant reflectance value
    - The surface is illuminated by a point light source infinitely far away





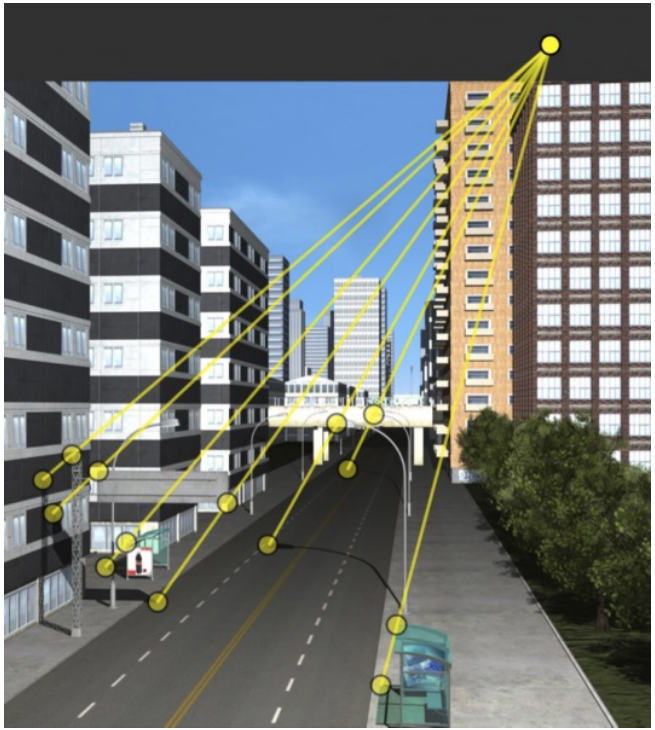


Detecting composite photo

Fake photo

Real photo

Shadow inconstistencies



Shadow inconstistencies

Highlights inconsistencies

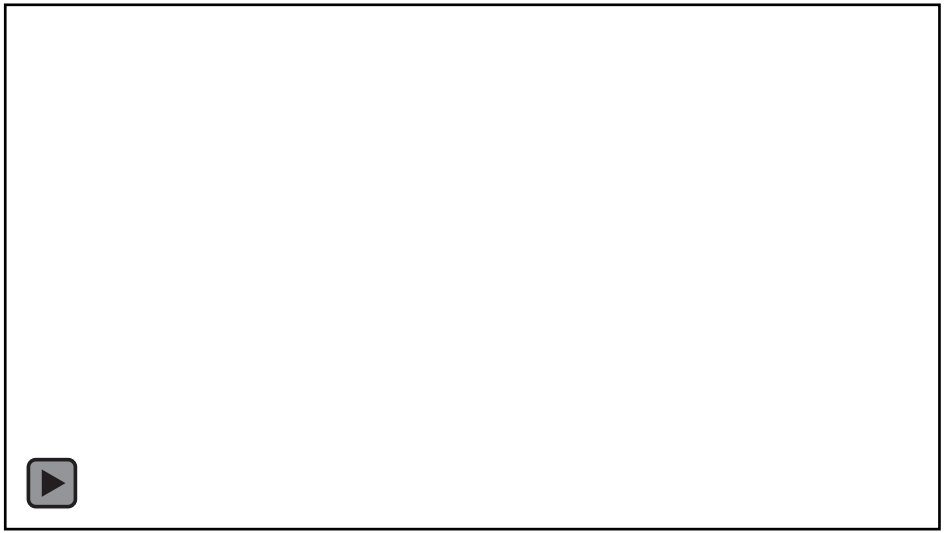
- The shape, color and position of eyes reflections can be used to determine if the lighting in a scene is consistent between two or more people.

Highlights inconsistencies

- The eyes are modelled as spheres and infer the direction of light source from highlights

Very likely this picture was taken in three different time instants

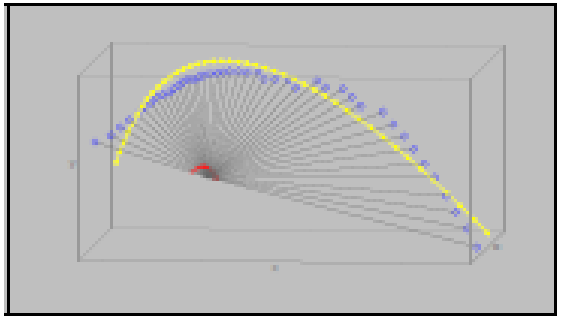
Trajectory inconsistencies (video)

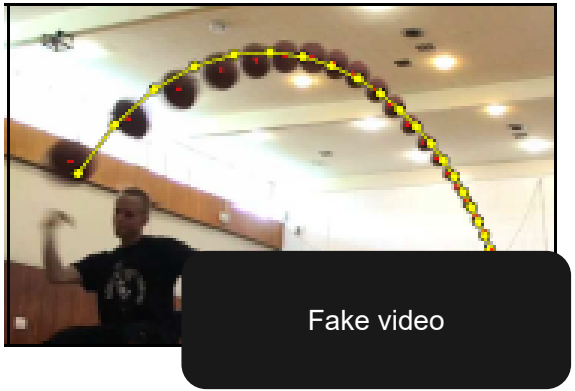


- https://www.youtube.com/watch?v=WbaH52JI3So

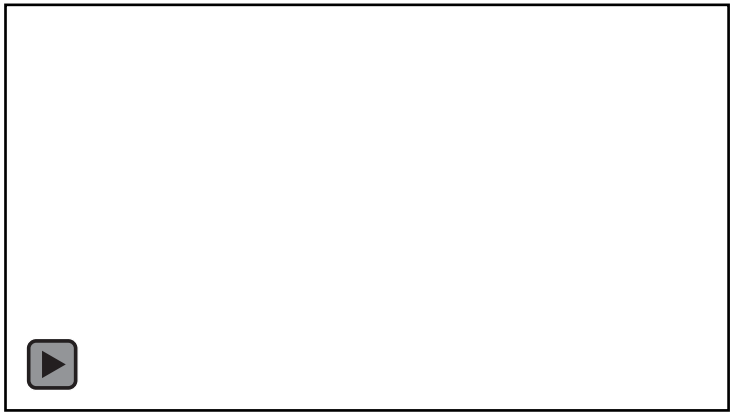
Trajectory inconsistencies (video)

- After perspective compensation we can reconstruct the 3D trajectory of the ball and compare it against the expected trajectory according to physics
- Comparing the expected and apparent trajectories we can deduce that the video is fake

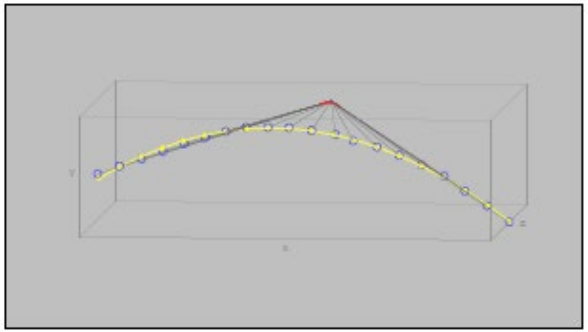


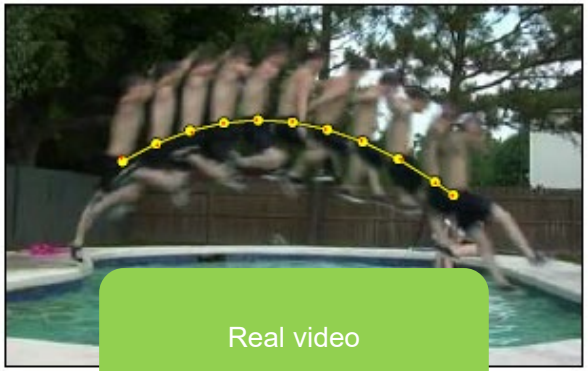


Another example



http://www.youtube.com/watch?v=wpYM3NkMYMs



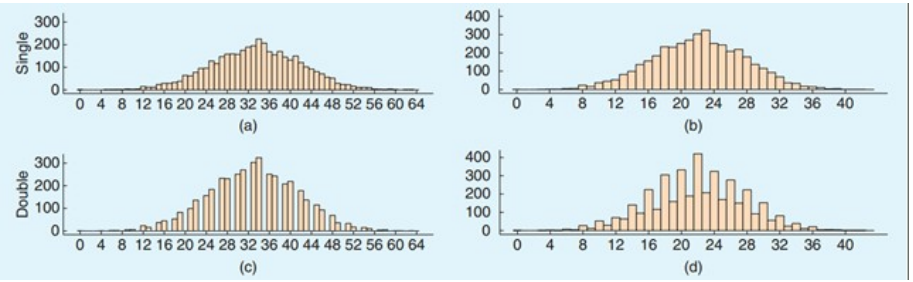


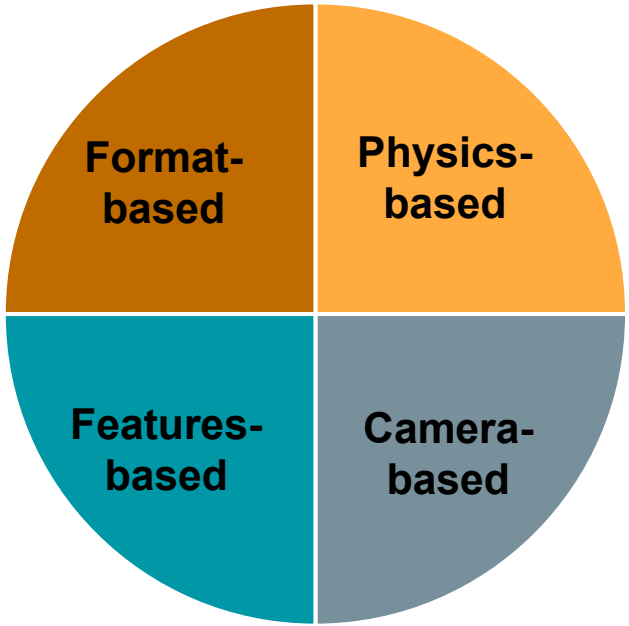
Real video

Techniques for tampering detection

- 4 categories

- Double JPEG compression
- First digits





Forgery creation

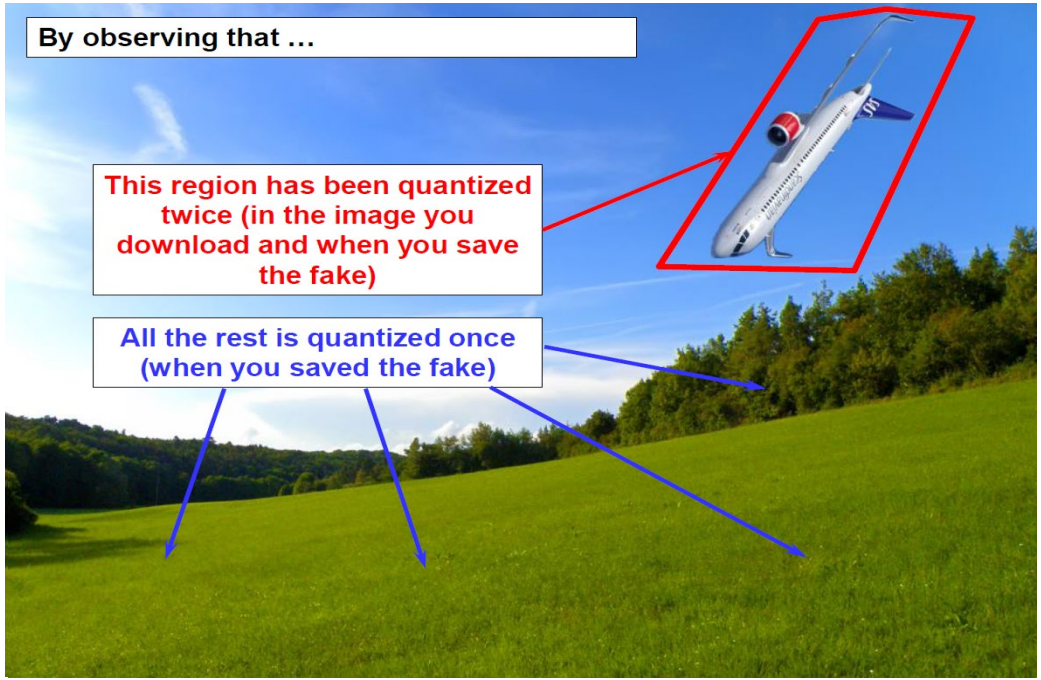
- Suppose you take a picture with your camera.
- Imagine that this picture has not undergone any compression (TIFF or RAW)

Forgery creation

- Download a JPEG image from the web.

Start your favorite image editing software ….

Forgery creation



Forgery detection with CNN

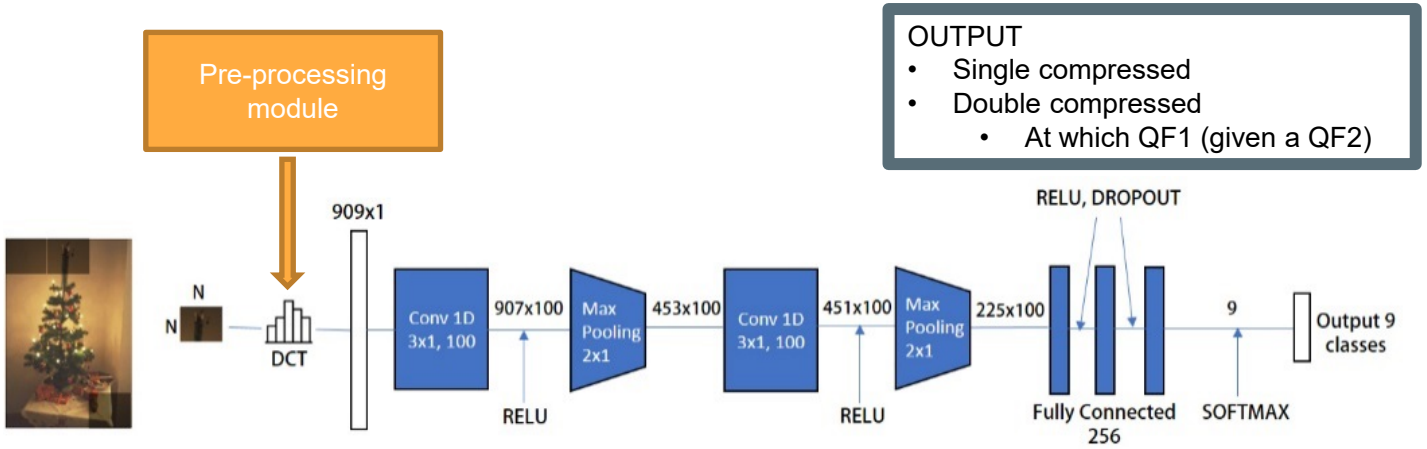
- Can a manipulation be revealed and localized with a Convolutional Neural Network?
- Find a universal method able to detect all the kind of attacks

I.Amerini, T. Uricchio, L. Ballan, R. Caldelli, "Localization of JPEG double compression through multi-domain convolutional neural networks", Media Forensics Workshop at CVPR 2017, July Honolulu, Hawaii 2017.

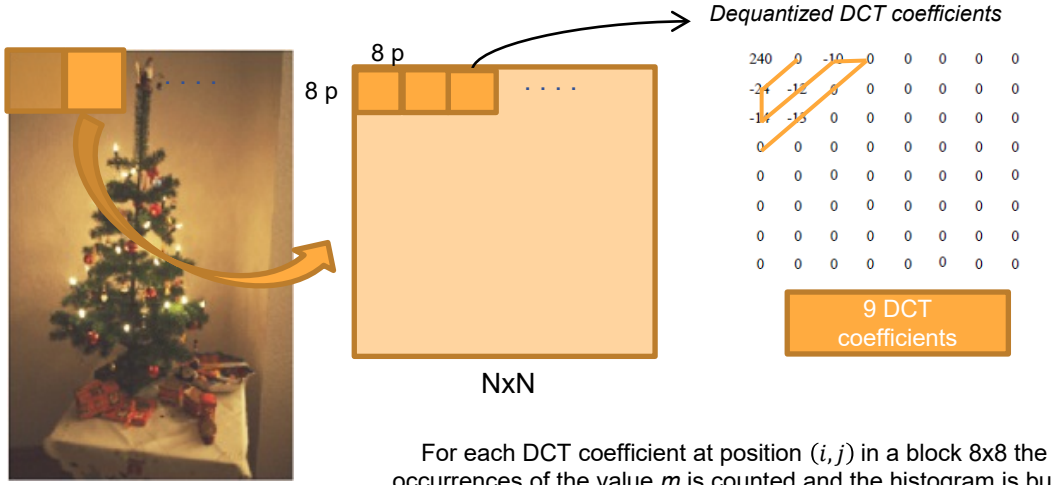
[WCVPR'17 Amerini et Al]

The proposed CNNs

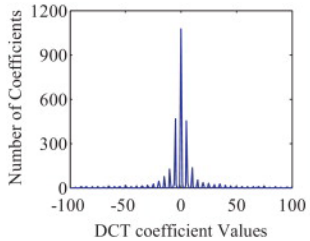
- Spatial-domain CNN
- Frequency-based CNN



Pre-processing module



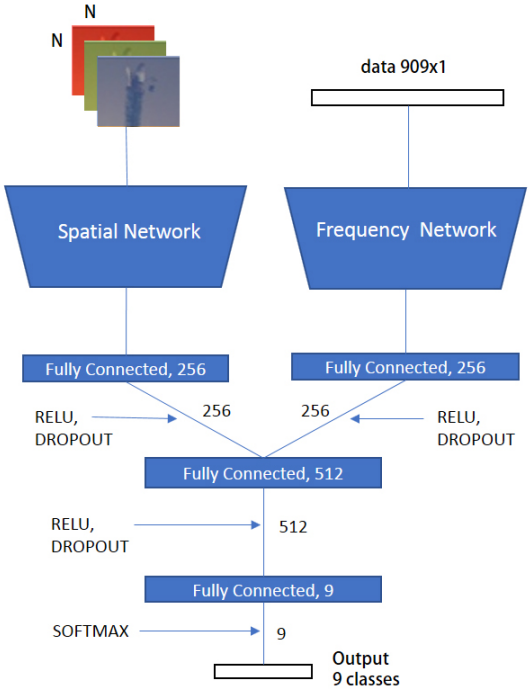
For each DCT coefficient at position (i, j) in a block 8x8 the occurrences of the value *m* is counted and the histogram is built.



$$h_{(i,j)}(m)$$
$$m = \{-50, 0, +50\}$$

Features=
[101 histogram bins x 9 coeff. DCT]

Multi-domain CNN

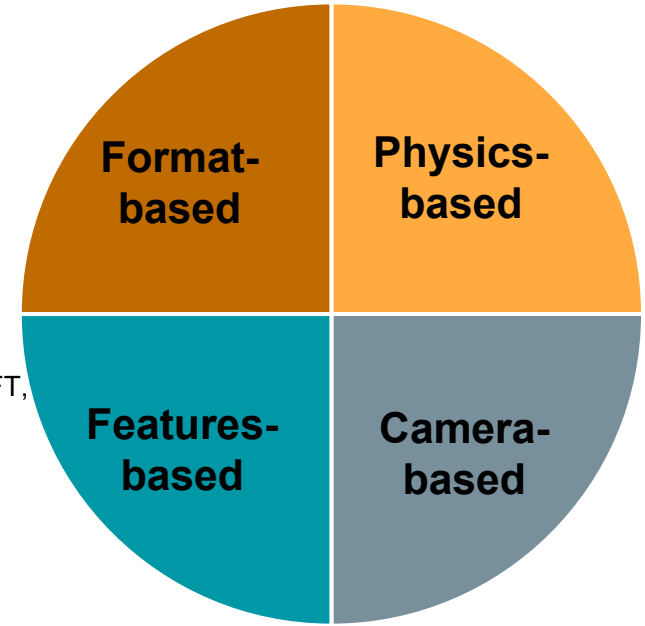


- The multi-domain CNN learns the intermodal relations between **RGB** features and **DCT** histograms
    - This net combines the output of the fully connected layers
    - The last fully connected layer has 512 elements
- Output 9 classes: single compression and 8 double compressions combinations

The *Ferguson* case

Techniques for tampering detection

- 4 categories

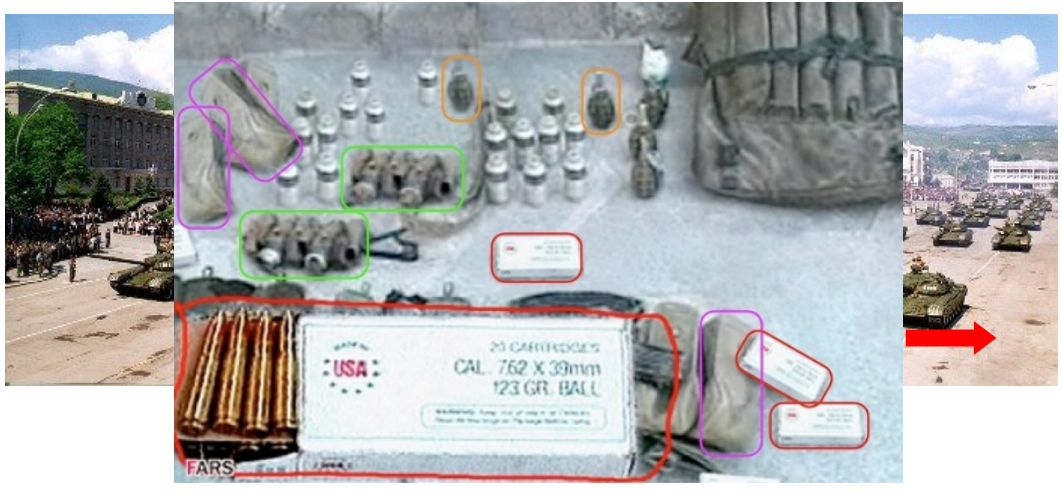


- Salient points (SIFT, SURF etc..)

The copy-move manipulation

Hiding something

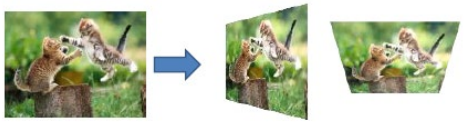
Duplicating something



Copy-Move Detection: salient point-based

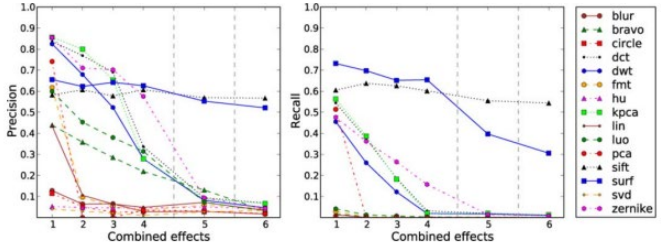
When performing a cloning, usually a geometric transformation is applied to the cloned patch.

TARGET:
Forensic analysis should provide instruments to detect such a cloning and to estimate which transformation has been performed.



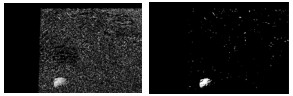
- In object detection and recognition, techniques based on scene modeling through a collection of salient points are well established.
- **SIFT** (*Scale Invariant Features Transform*) are usually adopted for their high performances and low complexity.

The proposed CMF detector



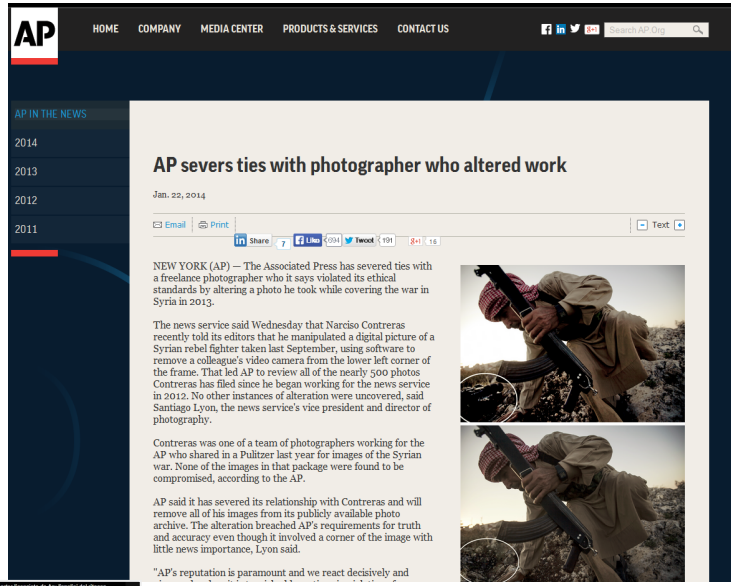
Scaling, rotation, JPEG compression [Riess, TIFS'12]

Geometric transformation estimation



Correlation mask and segmentation

The syrian soldier case



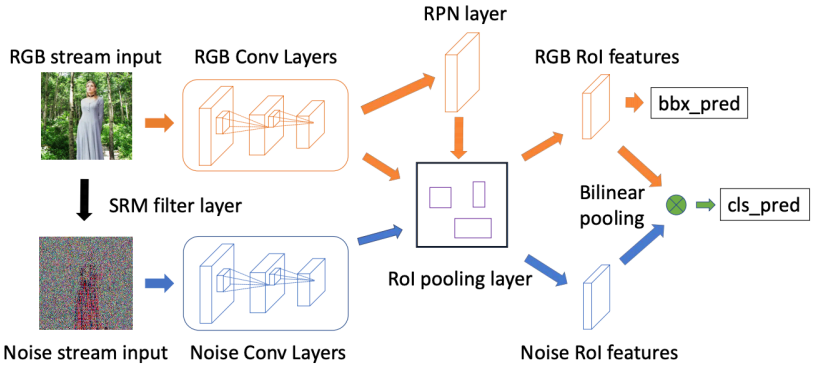
I. Amerini, et Al, "A SIFT–based forensic method for copy-move attack detection and transformation recovery". IEEE Transactions on Information Forensics and Security, 2011

Techniques for tampering detection

- Based on learned features

Two-stream Faster R-CNN (1/2)

- The CNN streams are two Faster R-CNN based on the ResNet101
- The RGB stream is used both for bounding box regression and manipulation classification
- The SRM filter (5×5×3) is used to pre-process the RGB image for the noise stream
- The feature outputs of the two streams are combined at each location using the matrix outer product called bilinear pooling
- A fully connected layer and a softmax layer output the predicted and determines whether predicted regions have been manipulated or not



Zhou et al., Learning Rich Features for Image Manipulation Detection, EEE/CVF Conference on Computer Vision and Pattern Recognition, 2018

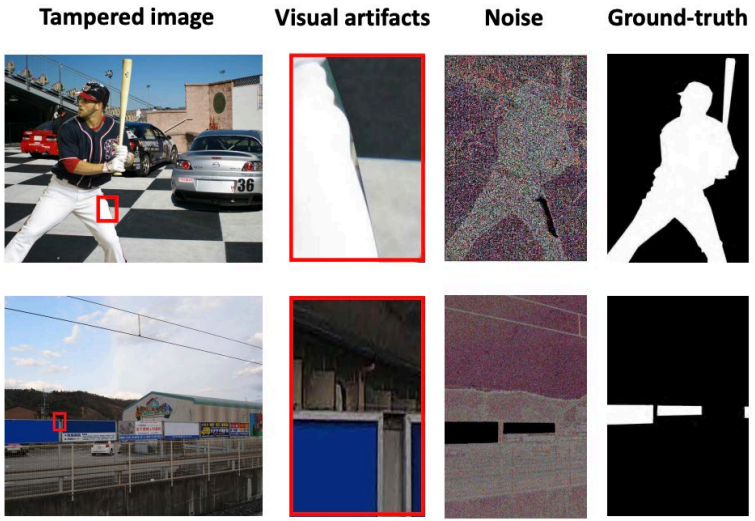
Two-stream Faster R-CNN (2/2)

- The network is trained end-to-end using cross entropy loss for manipulation classification and smooth L1 loss for bounding box regression

- They create a synthetic tampering dataset based on COCO for pre-training the model and then finetuned the model on different datasets for testing

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

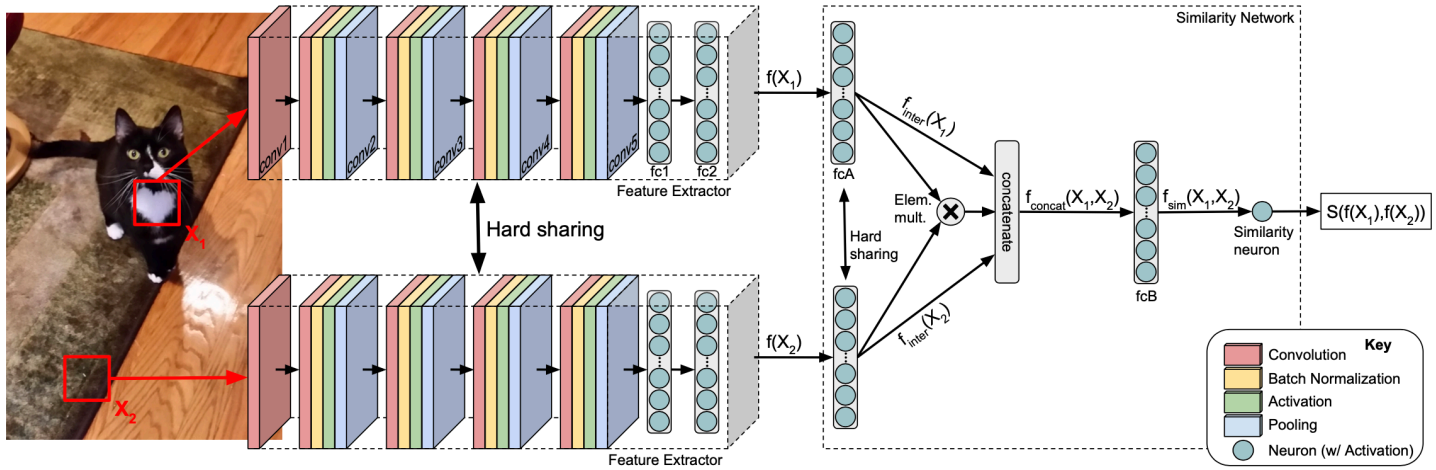
Figure 4. The three SRM filter kernels used to extract noise features.



Zhou et al., Learning Rich Features for Image Manipulation Detection, EEE/CVF Conference on Computer Vision and Pattern Recognition, 2018

Forensic similarity (1/2)

- Key idea: learn to measure the similarity (in terms of forensic features) of two images or patches
- The Siamese CNN-based feature extractors learn to extract high-level forensic feature vectors
- The similarity network is a neural network that maps feature vectors from two image patches to a similarity score indicating whether they contain the same or different forensic traces.

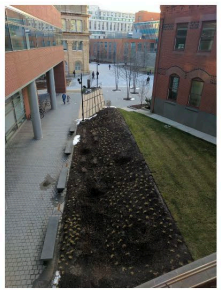


Mayer and Stamm, Forensic Similarity for Digital Images, IEEE Transactions on Information Forensics and Security 2020

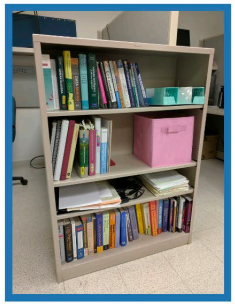
Forensic similarity (2/2)

Two learning phases:

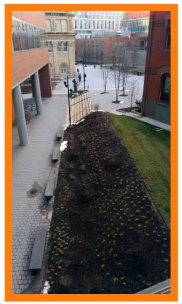
- **Phase A**. Add an additional fully-connected layer with softmax activation to the feature extractor architecture. Iteratively train the network using stochastic gradient descent with a cross-entropy loss function
- **Phase B**. Train the similarity network to learn a forensic similarity mapping for any type of measurable forensic trace, such as whether two image patches were captured by the same or different camera model or manipulated by the same or different editing operation. Allow the error to back propagate through the feature extractor and update the feature extractor weights.



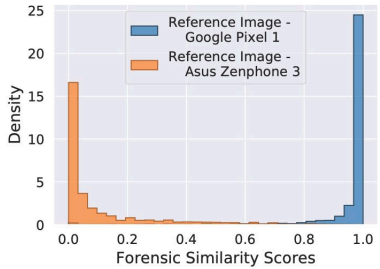
(a) Reference Image (Google Pixel 1)



(b) Google Pixel 1 Image

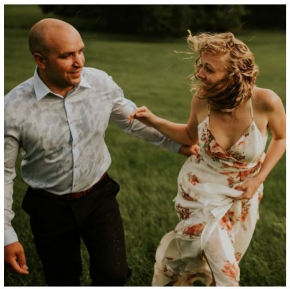


(c) Asus Zenfone 3 Image

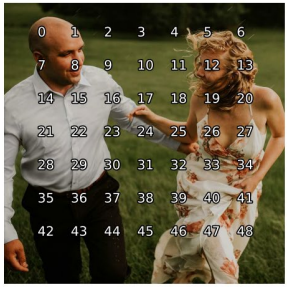


(d) Forensic Similarity Distribution

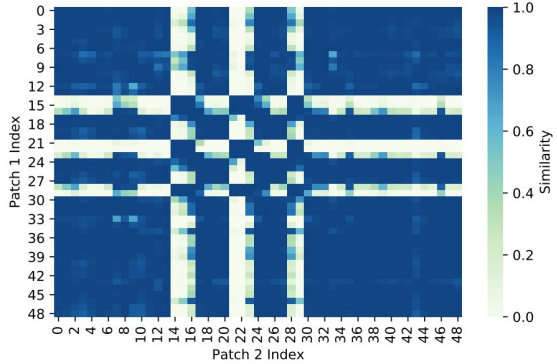
Example with forgery



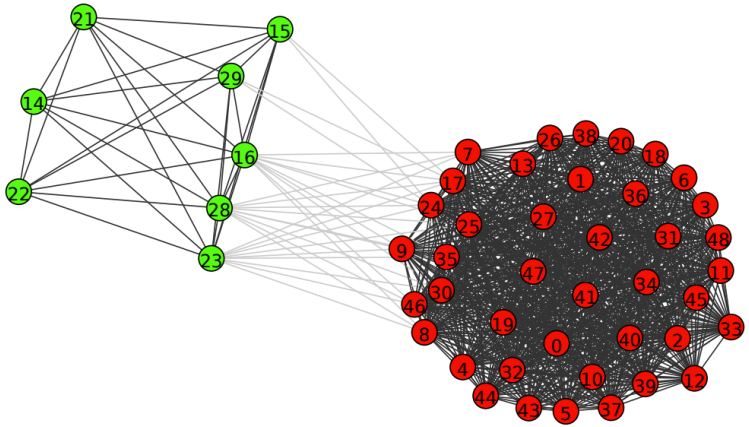
(a) Original Image



(b) Edited Image, with Patch Indices Overlaid



(c) Forensic Similarity Matrix



(d) Forensic Similarity Graph with Community Partitions

Adversarial deep learning

Counter Forensics/Adversarial Forensics

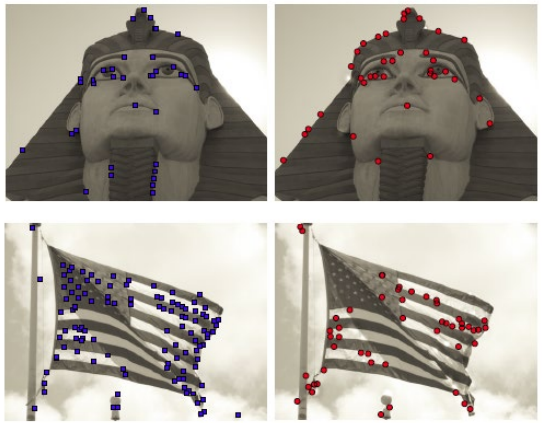
- The **forensic analyst** tries to recover the history of a digital image.
- The **adversary** employs counter-forensics methods to deceive forensic analyst.

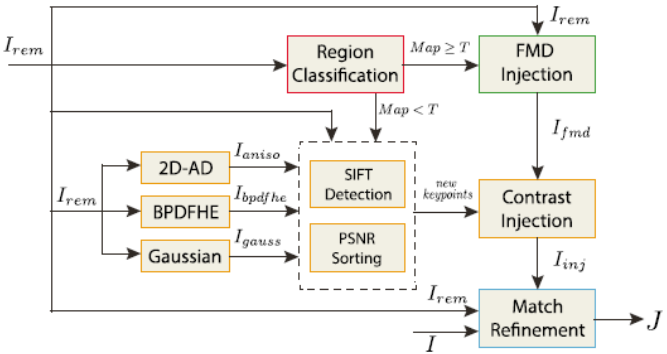
General idea:

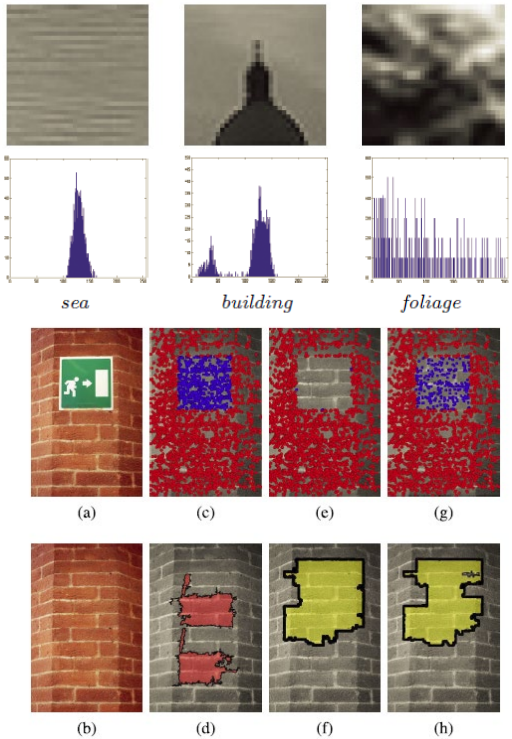
- If you know what statistic/method is used by the analyst
- Just adapt the statistic of your forgery to be very close to the statistic of good sequences
- Any detector based on the statistic will be fooled

Adversarial forensics on CMFD

- Robustness and security of the CMFD technique
  - SIFT keypoints removal and injection





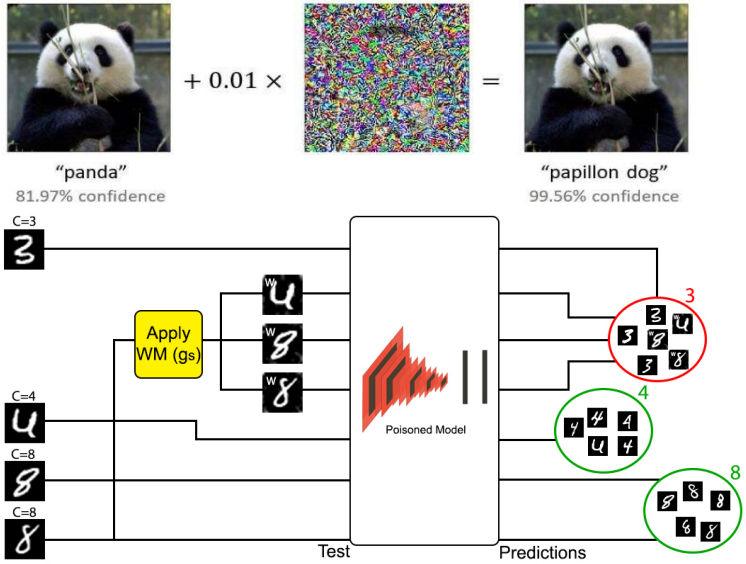


A. Costanzo, I. Amerini et Al, "Forensic Analysis of SIFT Keypoint Removal and Injection," Information Forensics and Security, IEEE Transactions on Information Forensics and Security, 2014

Adversarial deep learning

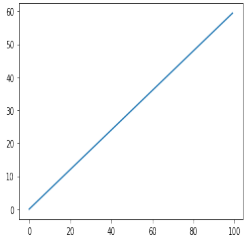
Adversarial machine learning, security and robustness of deep learning algorithms for computer vision sensible task

- Adversarial examples
  - At test time, evasion attacks
- Backdoor attacks
  - At training time (poisoning of training data for later exploitation)
  - Idea: **digital watermarking** to create a backdoor signal to fool a neural network in an image classification task

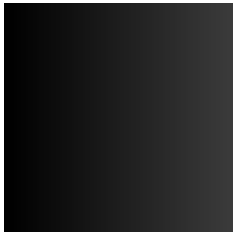


Caldelli, Abbate, Amerini, Exploiting Watermarking Algorithms to perform Backdoor Attacks in Convolutional Neural Networks, ICPR Workshop 2022

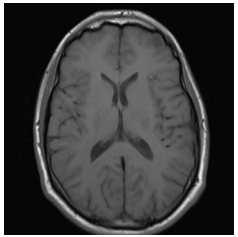
Backdoor Attack on MRI images



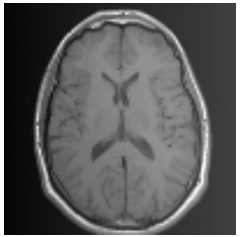
1 X 100
Ramp Signal



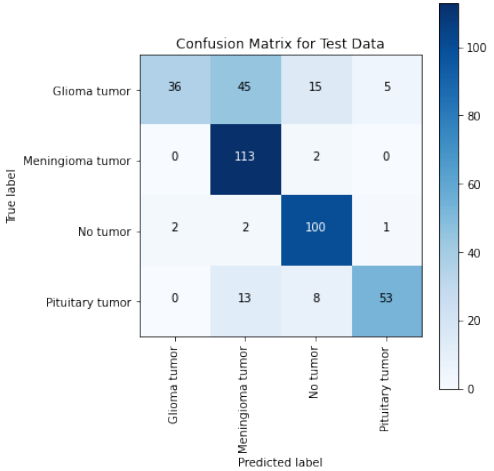
100 X 100
Ramp Signal

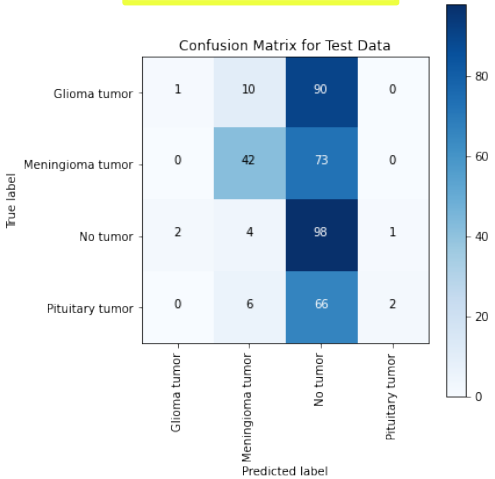


Normal Image



After adding ramp signal





ASR
79%

On going work - submitted

In general

- Fool a detector = force it to misclassify
- Approach: make the processed image statistic close to that of an untouched image

The cat and mouse game of Cyber Security!!!

Next week

- Deepfake detection methods
- GAN and diffusion model fingeprinting
- [Continual learning for fake news detection]