Fabrizio Guillaro

Towards Robust and General Image Forgery Detection and Localization

Tutor: Luisa Verdoliva

co-Tutor: Giovanni Poggi

Cycle: XXXVII

Year: Third

Candidate’s information

- **MSc degree** in Computer Engineering – Università degli Studi di Napoli Federico II
- **Research group**: GRIP (Image Processing Research Group)
- **PhD start date**: 01/11/2021
- **PhD end date**: 31/10/2024
- **Scholarship type**: funded by DARPA under the SEMAFOR program through the DISCOVER project
- **Periods abroad or in companies**:
  - 30/10/2023 - 29/01/2024 at Google LLC (Mountain View, California, USA)
  - 30/01/2024 - 10/05/2024 at Google S.r.l. (remotely in Italy)

Summary of study activities

PhD year	Courses	Seminars	Research	Tutorship
1st	26	10.8	23	1.28
2nd	14	4.1	41.1	0.28
3rd	13	0	47.4	0.5
Total	53	14.9	111.5	2.06

- **PhD Schools**:
  - *DeepLearn Summer School 2022* – Las Palmas de Gran Canaria, Spain
  - *International Computer Vision Summer School* (**ICVSS**) 2023 – Scicli (RG), Italy
  - *IEEE-EURASIP Summer School on Signal Processing* (**S3P**) 2024 – Capri (NA), Italy
- **PhD courses:**
  - *Introduction to Deep Learning* - Prof. Giovanni Poggi, Dr. Diego Gragnaniello
  - *How to boost your PhD* - Prof. Antigone Marino
  - *Statistical Multimedia Security and Forensics* - Prof. Fernando Pérez-González, at University of Trento
  - *Strategic Orientation for STEM Research & Writing* - Dr. Chie Shin Fraser
  - *Innovation and Entrepreneurship* - Prof. Pierluigi Rippa
- **MSc courses**:
  - *Visione per Sistemi Robotici* - Prof. Giovanni Poggi, Dr. Davide Cozzolino
  - *Image and Video Processing for Autonomous Driving* - Prof. Luisa Verdoliva
- **Conferences**:
  - *International Conference on Pattern Recognition* (**ICPR**), Montréal, Aug 21-25, 2022
  - IEEE *International Workshop on Information Forensics* (**WIFS**), (online) Dec 13-16, 2022
  - IEEE/CVF *Computer Vision and Pattern Recognition Conference* (**CVPR**), Vancouver, Jun 18-22, 2023

Research field of interest

- **Image Forensics**:
  - Analysis of forensic clues from visual data
- **Image Forgery Detection (IFD)**:
  - Is the image fake? Has the image been manipulated?
- **Image Forgery Localization (IFL)**:
  - Which part of the image has been manipulated?

Score

0.98

FAKE

itee PhD
information technology
electrical engineering

Research field of interest

- **Image Forensics**:
  - Analysis of forensic clues from visual data
- **Image Forgery Detection (IFD)**:
  - Is the image fake? Has the image been manipulated?
- **Image Forgery Localization (IFL)**:
  - Which part of the image has been manipulated?

Score

0.98

FAKE

Research field of interest

- **Image Forensics**:
  - Analysis of forensic clues from visual data
- **Image Forgery Detection (IFD)**:
  - Is the image fake? Has the image been manipulated?
- **Image Forgery Localization (IFL)**:
  - Which part of the image has been manipulated?
- **Synthetic Image Detection (SID)**:
  - Is the image generated by AI?

Real

AI generated

Research results

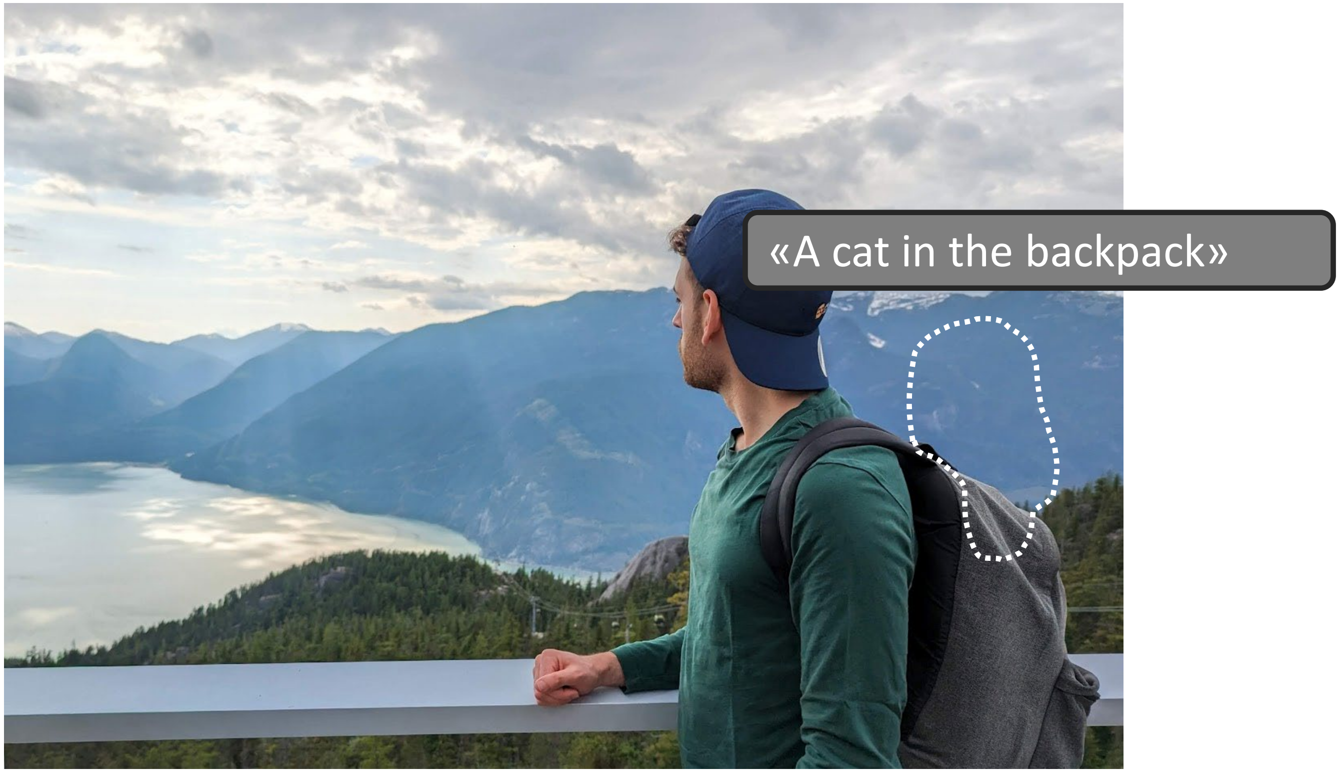
- Development of an IFL method (**Comprint**) based on the compression fingerprint of an image
- Development of a general IFL and IFD method (**TruFor**), based on:
  - A more robust noise fingerprint (Noiseprint++)
  - A confidence map for a more trustworthy detection
- Exploration of the **adversarial robustness** of Synthetic Image Detectors and transferability of the attacks

Research products

[P1]	H. Mareen, D. Vanden Bussche, F. Guillaro, D. Cozzolino, G. Van Wallendael, P. Lambert, L. Verdoliva, *Comprint: Image Forgery Detection and Localization using Compression Fingerprints*, **Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. Lecture Notes in Computer Science,** vol 13644, pp. 281-299. Springer, Cham. Montréal, QC, Canada, 2022
[P2]	F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Google Research *TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization*, **IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),** Vancouver, BC, Canada, 2023, pp. 20606-20615
[P3]	F. Guillaro, D. Cozzolino, G. Poggi, L. Verdoliva, *Uncertainty-driven detection and localization of image forgeries,* **Chapter in CNIT Volume. Series: Signal Processing and Learning for Next Generation Multimedia**, pp. 145-164, 2024
[P4]	V. De Rosa, F. Guillaro, G. Poggi, D. Cozzolino, L. Verdoliva, *Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection,* **IEEE International Workshop on Information Forensics and Security (WIFS),** Rome, Italy, December 2024.

PhD thesis: Overview

- Problem
  - **Editing tools** are easier to use and more powerful

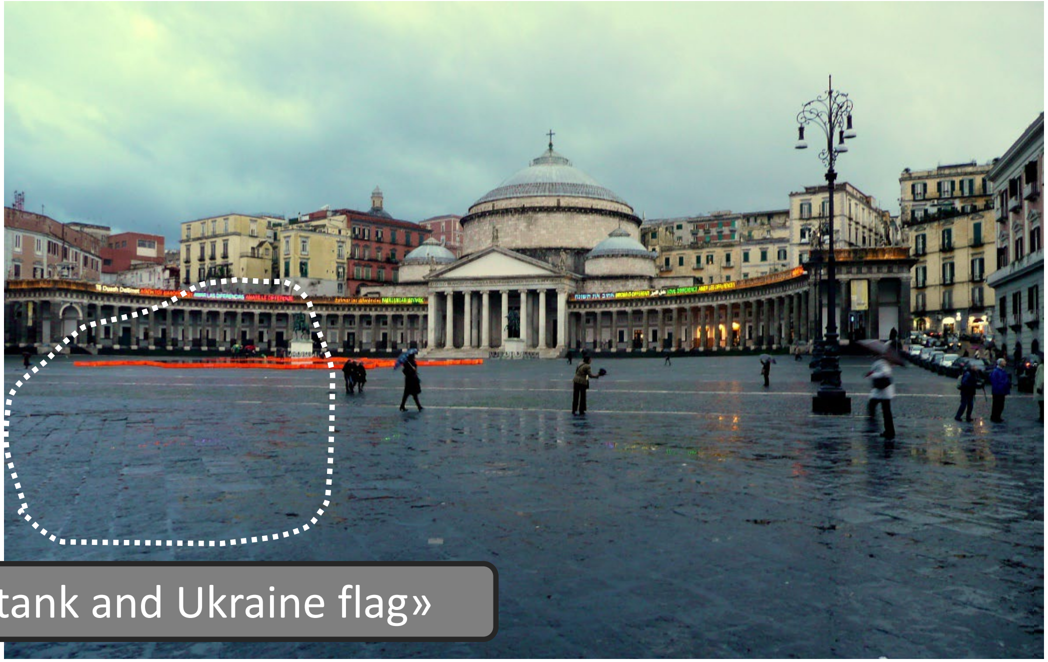


PhD thesis: Overview

- Problem
  - **Editing tools** are easier to use and more powerful

PhD thesis: Overview

- Problem
  - **Editing tools** are easier to use and more powerful
  - Users can maliciously manipulate data and spread **fake news**



«A tank and Ukraine flag»

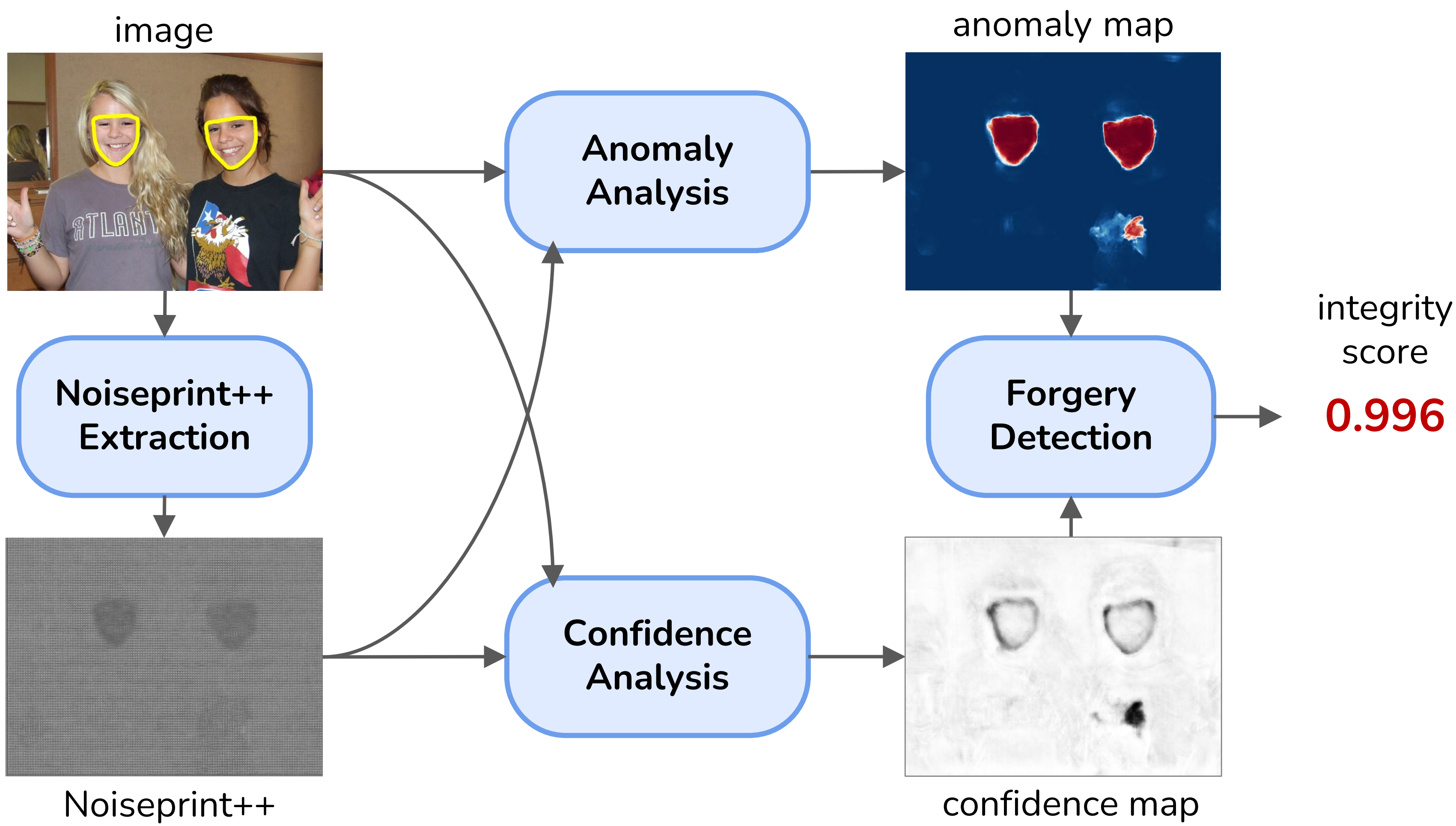
PhD thesis: Overview

- Problem
  - **Editing tools** are easier to use and more powerful
  - Users can maliciously manipulate data and spread **fake news**

PhD thesis: Overview

- Problem
  - **Editing tools** are easier to use and more powerful
  - Users can maliciously manipulate data and spread **fake news**
- Objective
  - Develop **general** techniques for image forgery **detection** and **localization**
  - Design methods that are **robust** to post-processing operations, such as re-compression and resizing

TruFor: Overview



Phase

1

Phase

2

Phase

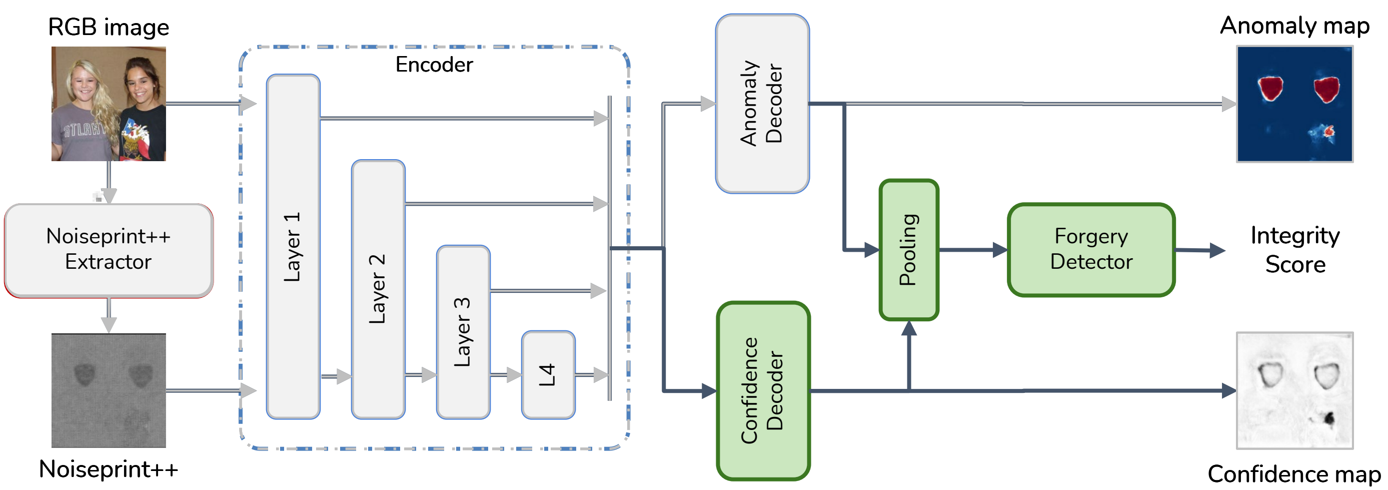
3

Methodology

1 Noiseprint++ Extraction
- A **noise-sensitive fingerprint** with high-level information
- Training: only pristine images

2 Anomaly Localization
- **Cross-modal** framework (RGB and NP++)
- Training: pristine and forged images

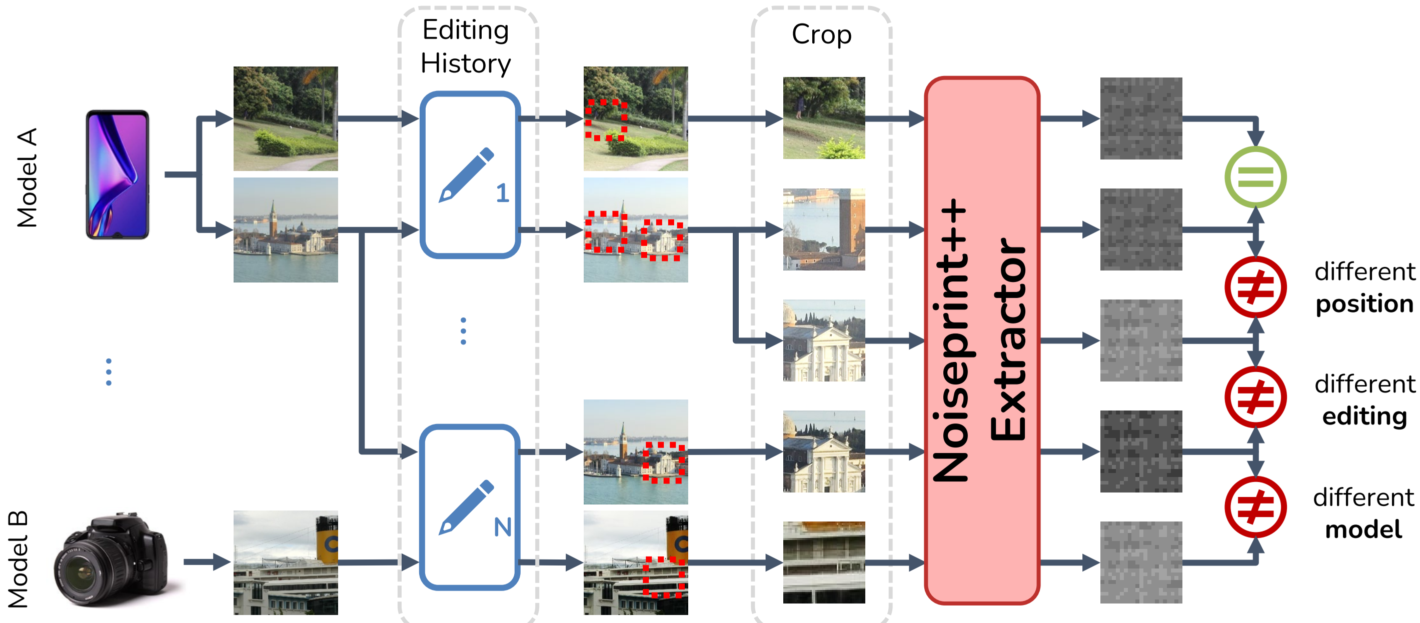
3 Confidence Estimation and Forgery Detection
- Confidence and anomaly maps for a **reliable** detection
- Training: pristine and forged images



Noiseprint++ extractor

- Contrastive Learning only on **real images** (**to gain generalization**)
- Training includes around 25k images from 1500 camera models

  (8 patches per image with random editing history)



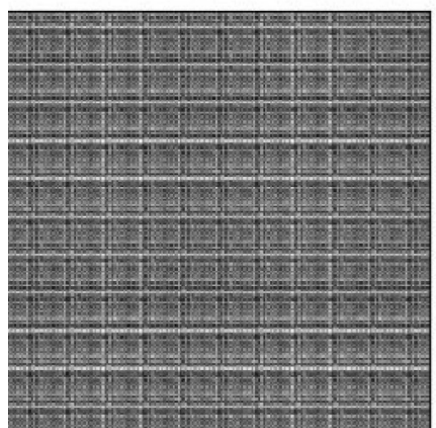
itee PhD information technology electrical engineering
Google Research

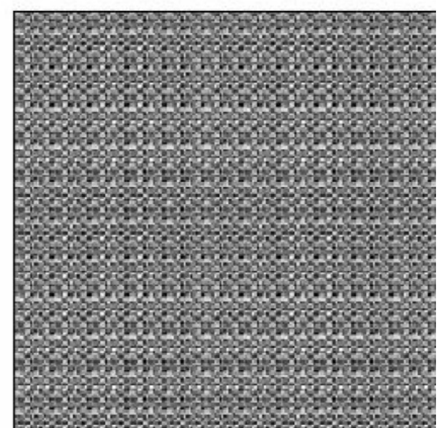
Noiseprint++

- It’s a **learned noise residual,** which enhances high frequency traces and suppresses the semantic content
- A first attempt was made with *Comprint*, a compression fingerprint which only enhanced JPEG compression artifacts
- *Noiseprint++*, instead, represents a fingerprint of the camera model and **editing history** of an image (**to gain robustness**)

Combination of:
resizing,
compression,
adjustments

Samsung Galaxy S3 Mini



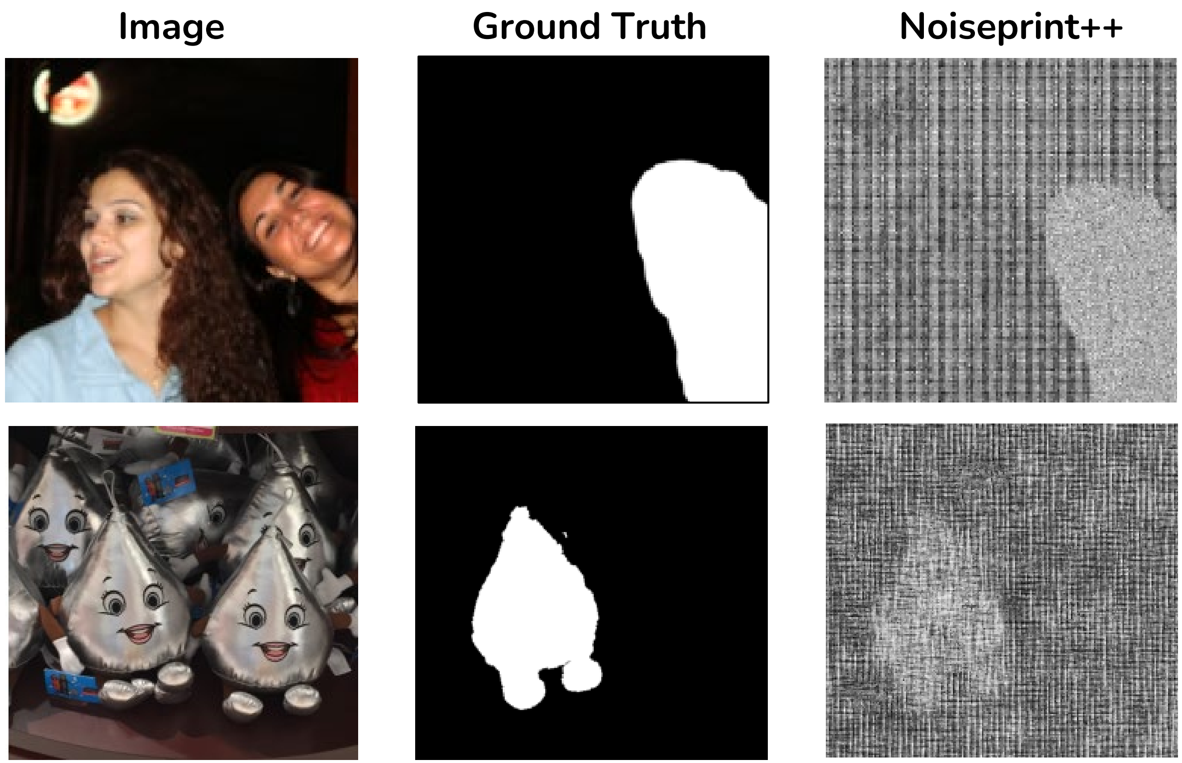


Apple iPhone 5c

itee PhD information technology electrical engineering

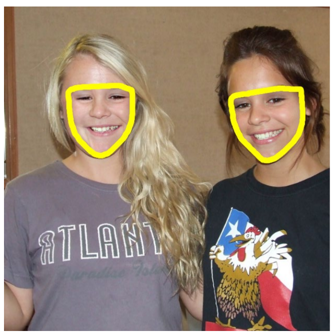
Noiseprint++

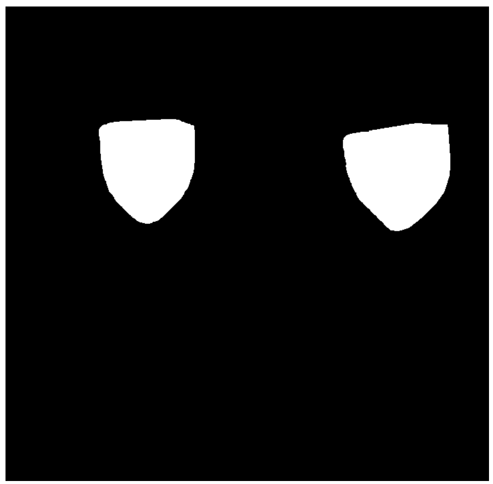
- When an image is manipulated, the noise pattern is disrupted
- **Inconsistencies** between forged and pristine regions are enhanced with improved robustness to post-processing operations



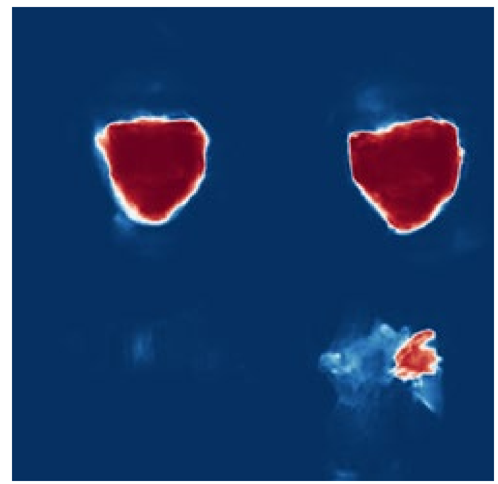
Confidence estimation

- Anomaly localization maps may have **false positives**
- We develop a strategy that estimates a pixel-level **confidence map**





ground truth (gt)



anomaly map ($pred$)

Confidence estimation

- Our confidence criterion is **True Class Probability** (TCP):
  - for each pixel it is the value corresponding to the true class

$$TCP_i = gt_i \cdot pred_i + \ (1 - gt_i)(1 - pred_i)$$

Confidence (TCP)

good prediction → **1** (white)

bad prediction → **0** (black)

Confidence estimation

- Ground truth is needed for TCP, but we do not have it at inference time
- We need to estimate it with a **learned confidence**

estimated confidence

Confidence estimation

- Ground truth is needed for TCP, but we don't have it at inference time
- We need to estimate it with a **learned confidence**

Confidence (TCP)

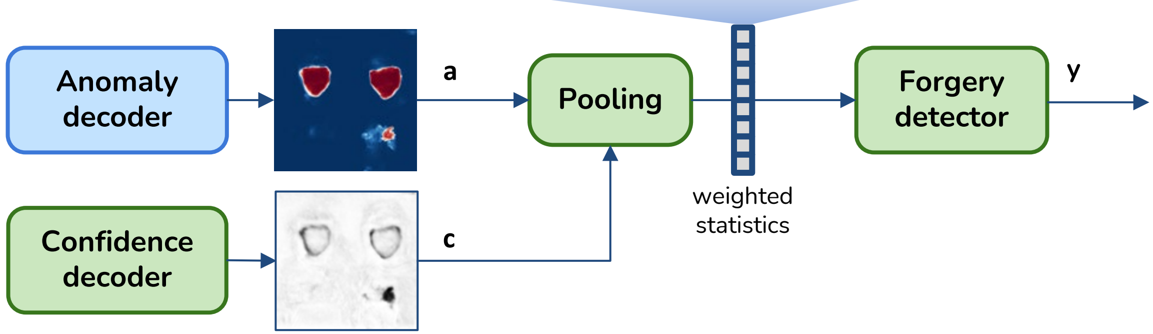
estimated confidence

Forgery detection

- **Confidence estimation** and the **detector** networks trained together
- Eight statistics are fed to the detector

$$c_{\text{avg}} = \sum_i c_i \qquad c_{\text{msq}} = \sum_i c_i^2 \qquad c_{\text{min}} = -\log \sum_i e^{-c_i} \qquad c_{\text{max}} = \log \sum_i e^{c_i}$$

$$a_{\text{avg}} = \sum_i \acute{c}_i\, a_i \qquad a_{\text{msq}} = \sum_i \acute{c}_i\, a_i^2 \qquad a_{\text{min}} = -\log \sum_i \acute{c}_i\, e^{-a_i} \qquad a_{\text{max}} = \log \sum_i \acute{c}_i\, e^{a_i}$$



Forgery detection

- False positives in the localization map do not affect the final score
- A score > 0.5 indicates manipulation

image | **localization map** | **confidence map** | **integrity score** (is it fake?)

FAKE

~~0,837~~ (fake) (fake)

0,996

REAL

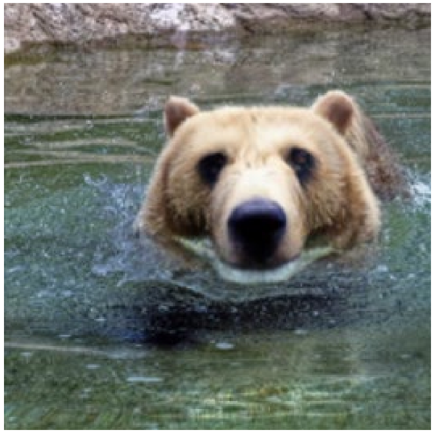
~~0,613~~ (fake) (real)

0,264

Metrics

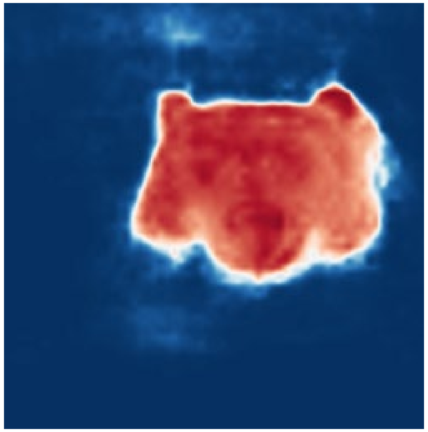
- Evaluation metrics:
  - Pixel-level localization metric: $F1 = \left(\frac{1}{precision} + \frac{1}{recall}\right)^{-1}$
  - Image-level detection metric: *Area Under ROC Curve*

image



ground truth

localization map



F1 score: 0.82

integrity score
(is it fake?)

0,93 (fake)

Evaluation results - Localization

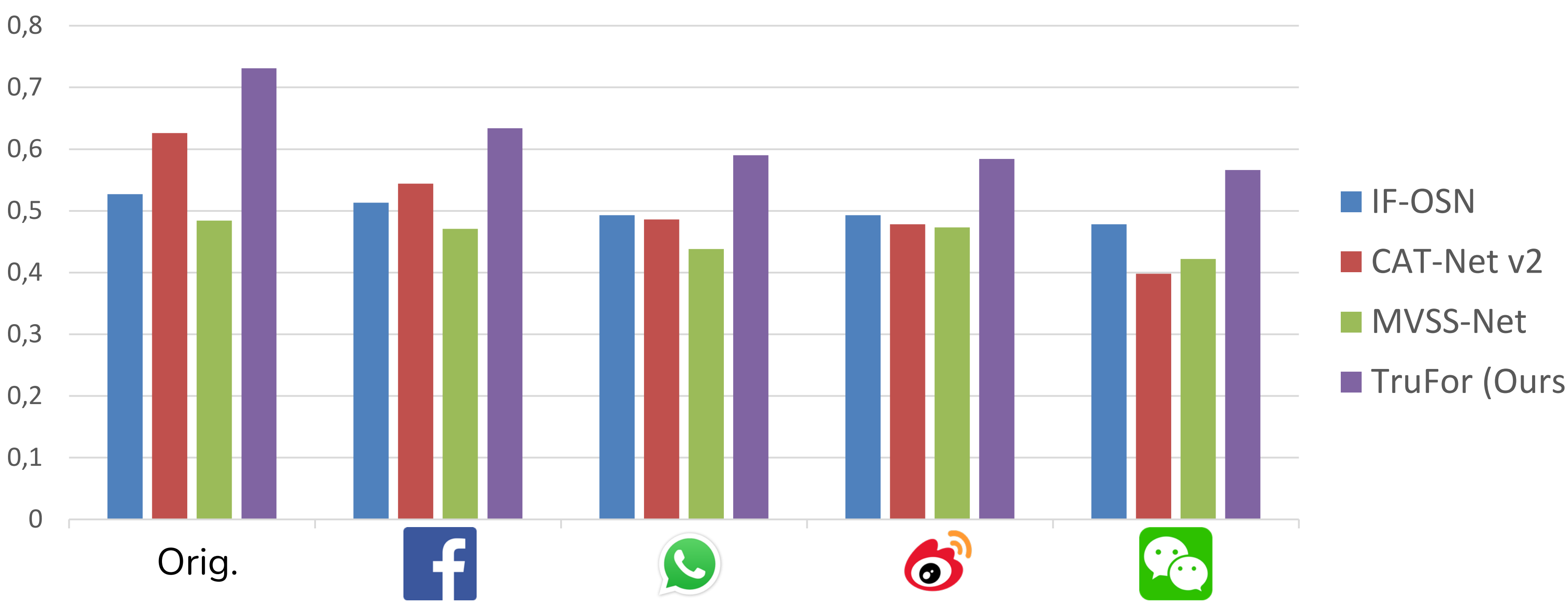
- Evaluation in terms of F1 on 8 publicly available datasets (4K fake images)
- Our method provides the best performance and it is able to generalize better

Method	CASIAv1	Coverage	Columbia	NIST16	DSO-1	VIPP	OpenFor	CocoGlide	AVG
Splicebuster	.252	.321	.811	.312	.662	.432	.459	.434	.460
EXIF-SC	.255	.332	.880	.298	.577	.424	.318	.424	.437
CR-CNN	.538	.487	.779	.363	.377	.355	.143	.577	.452
ManTraNet	.320	.486	.650	.225	.537	.373	.661	.673	.491
SPAN	.169	.428	.873	.363	.390	.375	.176	.350	.391
CAT-Net v2	.852	.582	.923	.417	.673	.672	.947	.603	.709
IF-OSN	.676	.472	.836	.449	.621	.508	.204	.589	.544
MVSS-Net	.650	.659	.781	.372	.459	.485	.225	.642	.534
PSCC-Net	.670	.615	.760	.210	.733	.309	.353	.685	.542
Noiseprint	.205	.342	.835	.345	.811	.546	.675	.405	.521
TruFor (Ours)	.822	.735	.914	.470	.973	.746	.901	.720	**.785**

Pixel-level F1
using best threshold per image

Robustness analysis

- Evaluation results on datasets uploaded on different **social media**
- When images are uploaded on the web, they undergo post-processing operations (resizing, compression, …)



Pixel-level F1
using fixed threshold

Evaluation results - Detection

- Evaluation in terms of AUC (0.5 represents the random guessing)
- Thanks to the use of the confidence map, TruFor performs better

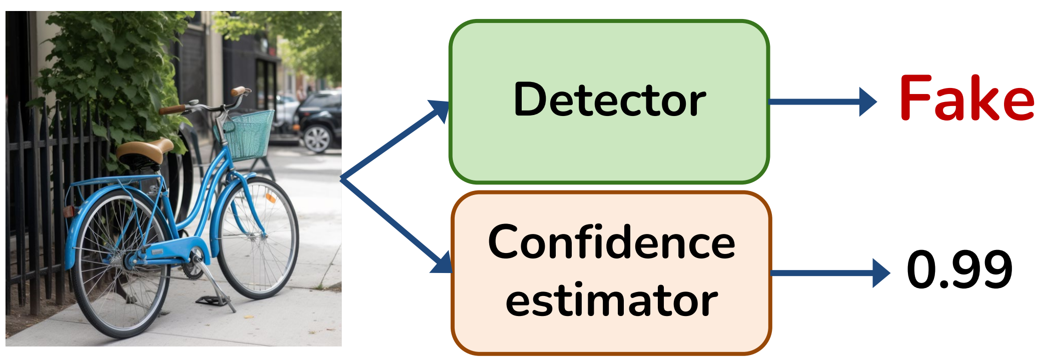
Method	CASIAv1+	Coverage	Columbia	NIST16	DSO-1	VIPP	CocoGlide	AVG
Splicebuster	.406	.541	.597	.610	.751	.539	.529	.568
EXIF-SC	.490	.498	.976	.504	.764	.617	.526	.625
CR-CNN	.670	.553	.755	.737	.576	.504	.589	.626
ManTraNet	.644	.760	.810	.624	.874	.530	.778	.717
SPAN	.480	.670	.999	.632	.669	.580	.475	.644
CAT-Net v2	.942	.680	.977	.750	.747	.813	.667	.797
IF-OSN	.735	.557	.882	.658	.853	.696	.611	.713
MVSS-Net	.932	.733	.984	.579	.552	.629	.654	.723
PSCC-Net	.869	.657	.300	.485	.650	.574	.777	.616
E2E	.377	.494	.894	.718	.803	.617	.530	.633
Noiseprint	.494	.525	.872	.618	.821	.580	.520	.633
TruFor (Ours)	.916	.770	.996	.760	.984	.820	.752	**.857**

Image-level AUC

Synthetic Image Detectors

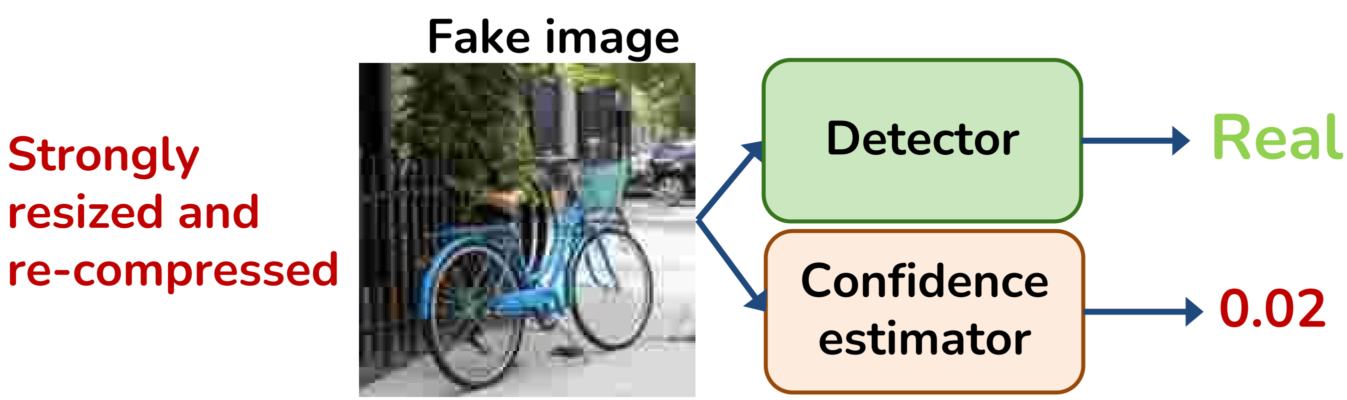
- We extend the idea of confidence estimation for the detection of fully AI-generated images
- This can help to discard the prediction if the detector is not confident enough (heavy post-processing)

Fake image



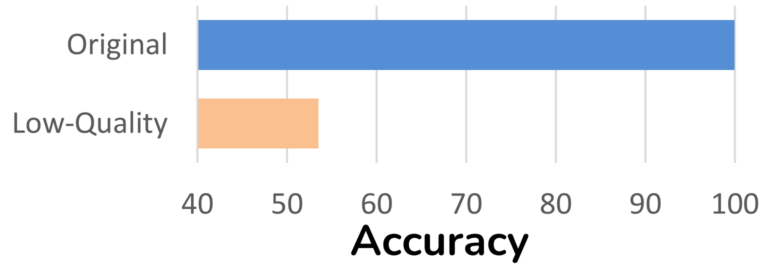
Synthetic Image Detectors

- We extend the idea of confidence estimation for the detection of fully AI-generated images
- This can help to discard the prediction if the detector is not confident enough (heavy post-processing)

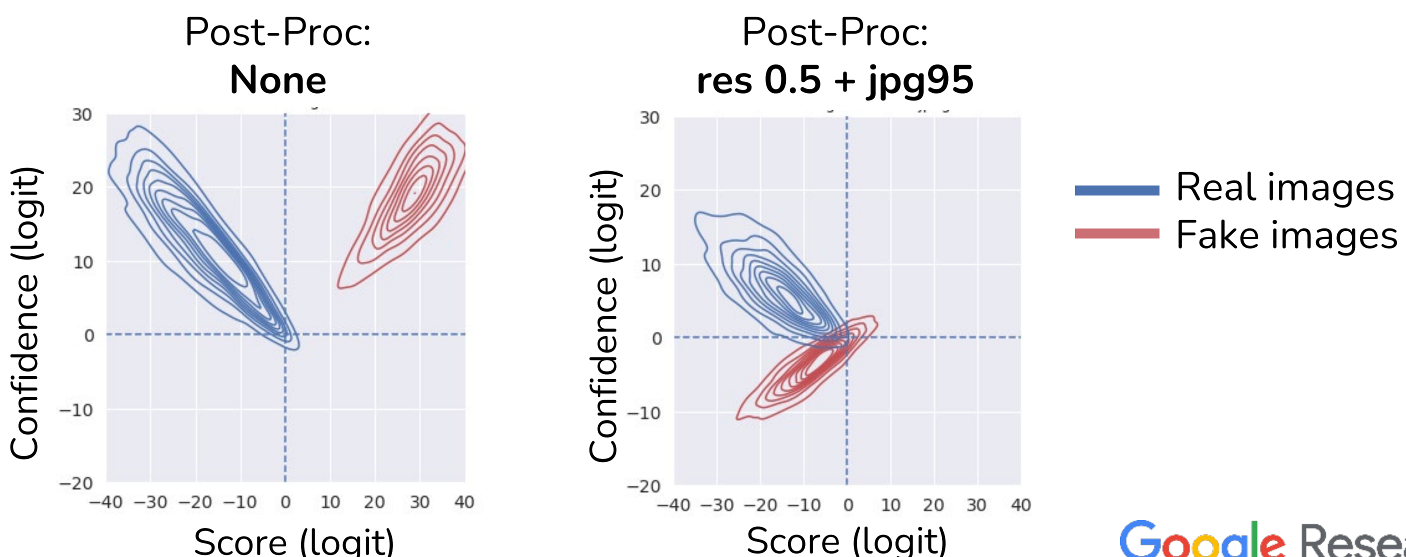


Synthetic Image Detectors

- The accuracy on low-quality data drops to less than 60%



- Fakes classified as real (red distribution leaning to the left) are marked as **unreliable** (distribution falls in the bottom of the graph)



itee PhD information technology electrical engineering

Google Research

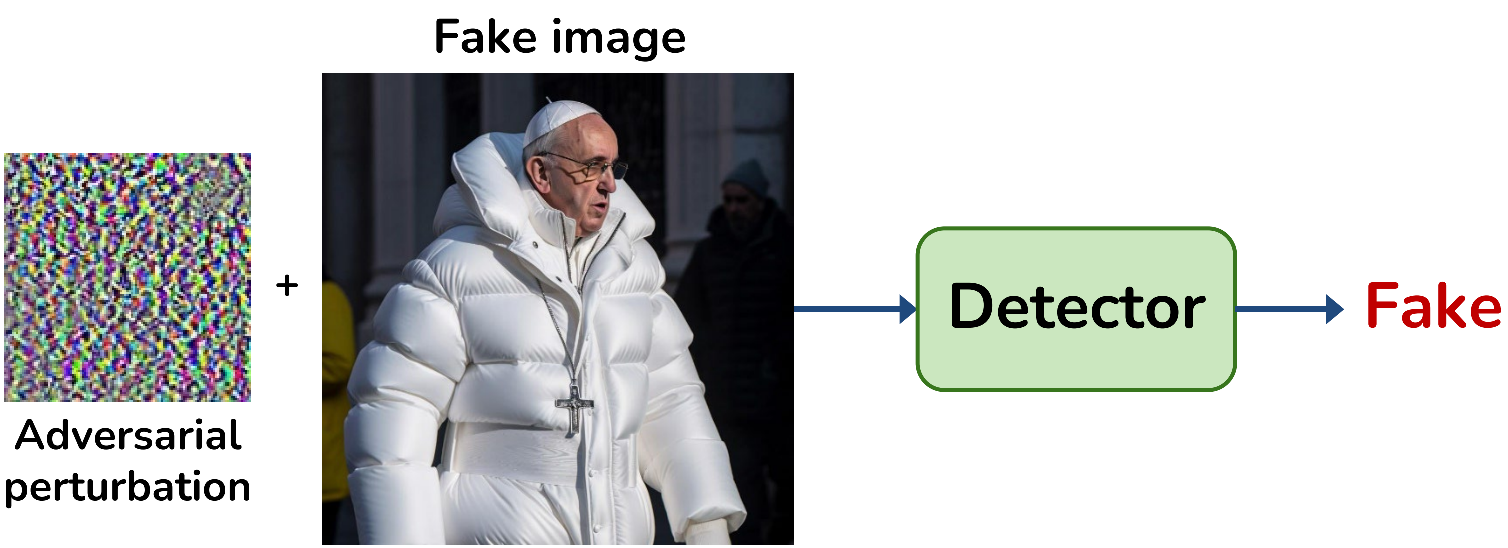
Adversarial Attacks

- An **adversarial attack** is designed to fool a detector into predicting a wrong label
- The attacked image is perturbed with an adversarial noise imperceptible to the naked eye

Fake image

Adversarial Attacks

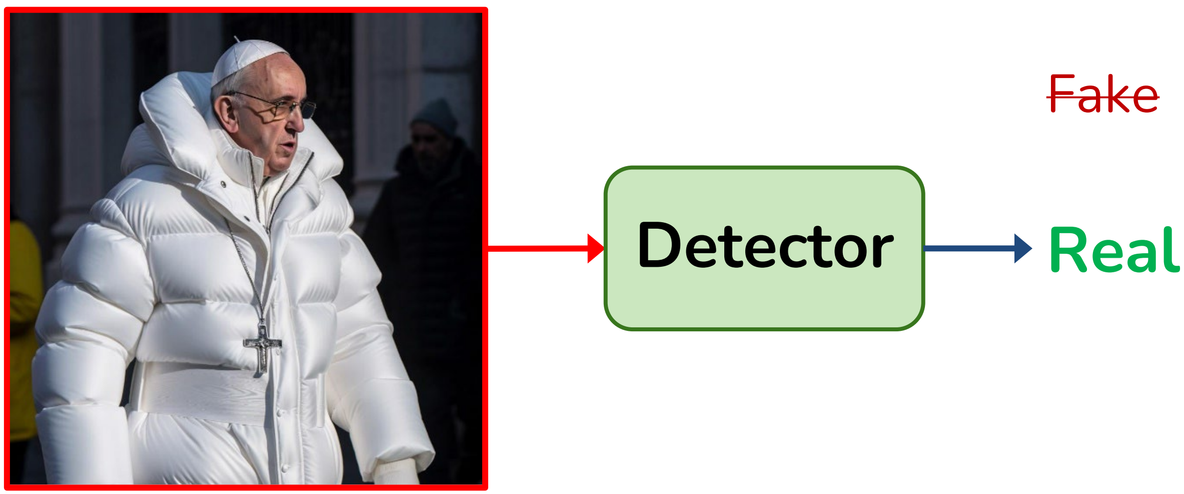
- An **adversarial attack** is designed to fool a detector into predicting a wrong label
- The attacked image is perturbed with an adversarial noise imperceptible to the naked eye



Adversarial Attacks

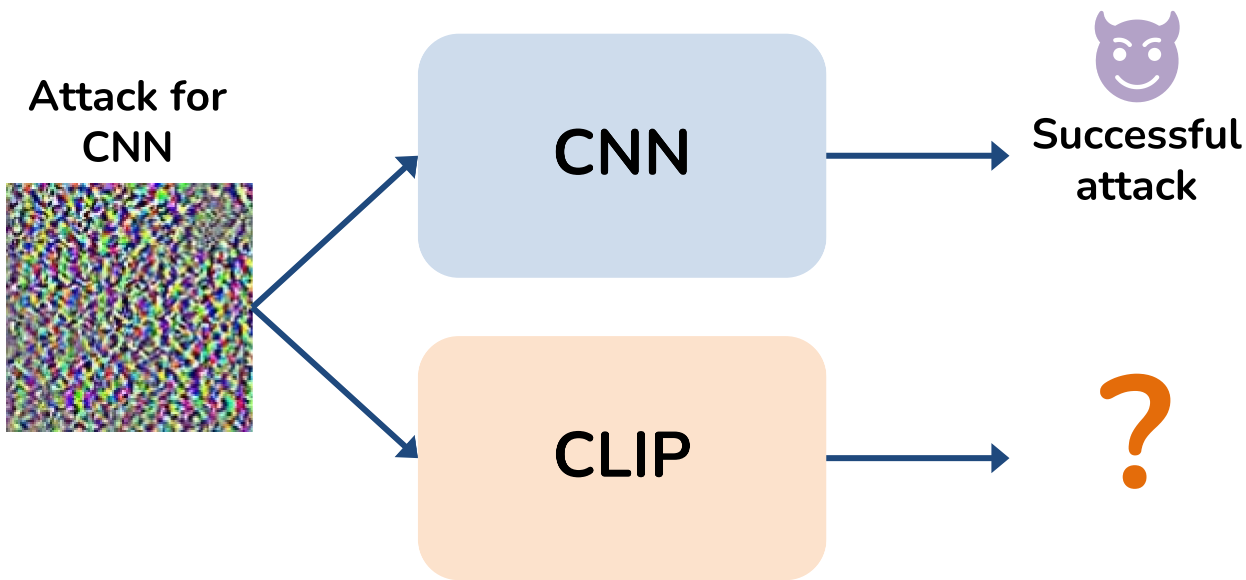
- An **adversarial attack** is designed to fool a detector into predicting a wrong label
- The attacked image is perturbed with an adversarial noise imperceptible to the naked eye

Attacked fake image



Adversarial robustness

- We explored the **adversarial robustness** of Synthetic Image Detectors to different attacks (l_2-PGD, DI^2-FGSM, RWA, UA)
- We analyzed the **transferability** of attacks between families of detectors
  - **CNN-based** (Convolutional Neural Networks)
  - **CLIP-based** (Contrastive Language-Image Pretraining)

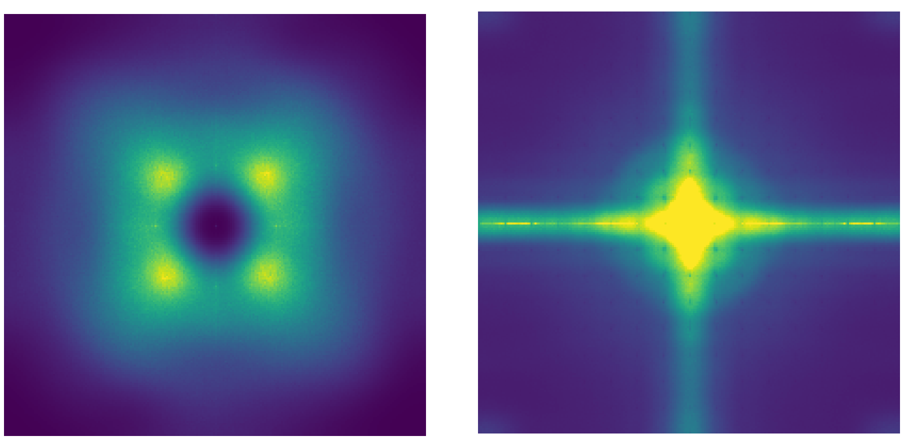


Results

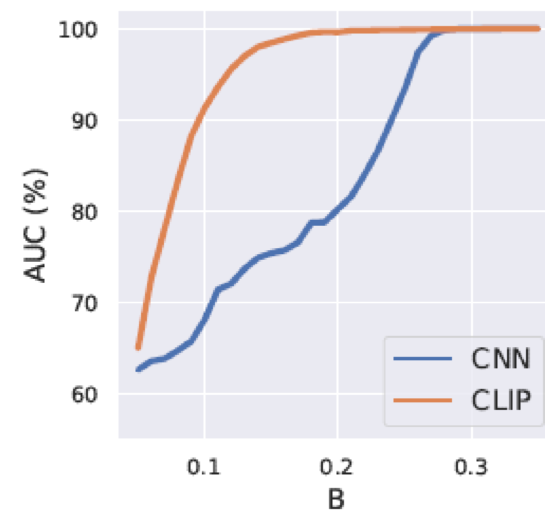
- **Findings**:
  - Attacks transfer easily between similar architectures…
  - …but **do not** transfer well between different families (CNN vs CLIP)

- **Explanation**:
  - CNN and CLIP detectors look at different frequencies

CNN CLIP



Power spectra
of adversarial noise patterns



Performance
in function of bandwith

Conclusions

- We introduced a general and robust Image Forgery Localization and Detection method based on contrastive learning and confidence map estimation
- We explored the adversarial robustness of Synthetic Image Detectors and transferability of attacks, shedding light on how forensic detectors work
- This analysis can help to build more effective detectors, robust to post-processing operations and to malicious attackers
- It would be also important to develop a strategy to detect both local and fully generated AI-content at the same time

Thank you for the attention!