UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Giada Zingarini

Towards General Integrity Verification Approaches for Natural and Medical Images

Tutor: Prof. Luisa Verdoliva

Cycle: XXXVIII Year: Third

Candidate's information

- **MSc degree** in Biomedical Engineering – Università degli Studi di Napoli Federico II
- **Research group:** GRIP (Image Processing Research Group)
- **PhD start date**: 01/11/2022
- **PhD end date:** 31/10/2025
- **Scholarship type:** funded by DARPA under the SEMAFOR program through the DISCOVER project
- **Periods abroad:**

  - 01/05/2025 – 28/07/205 at Recod.ai Lab (University of Campinas, Brazil)

Summary of study activities

PhD year	Courses	Seminars	Research	Tutorship
1st	26	5.3	30.8	0.00
2nd	9	5.4	41.0	0.52
3rd	5.2	0.4	57.0	0.00
Total	40.2	11.1	128,8	0.52

- **PhD Schools**:
  - IEEE-SPS Summer School 2023 “Summer School on Metaverse Technologies” – Cagliari (CA), Italia
  - IEEE-SPS Summer School 2024 “Understanding and modeling the world around us” – Capri (NA), Italia
  - IEEE-SPS Summer School 2025 “From Foundational Models to Multimedia Signal Processing: A deep dive in Multimodal AI” – San Vincenzo (LI), Italia
- **PhD courses**:
  - Using Deep Learning Properly – Dr. Andrea Apicella
  - *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
- **MSc courses**:
  - *Elaborazione di Segnali Multimediali* - Prof. Luisa Verdoliva
- **Conferences**:
  - IEEE *International Workshop on Information Forensics* (**WIFS**), (online) Dec. 13-16, 2022
  - IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP**), Seoul, South Korea, Apr. 14-19, 2024

Research field of interest

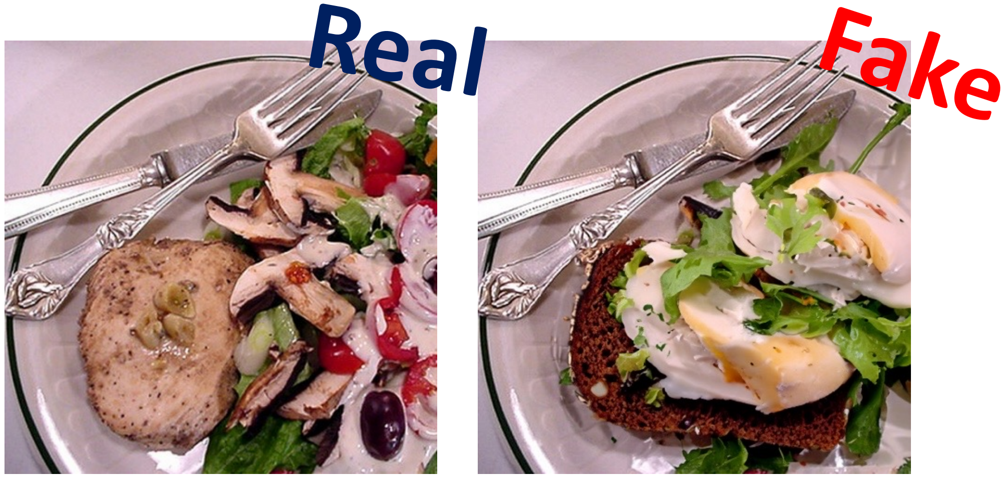
- **Multimedia Forensics**:
  - Analysis of multimedia data for forensic applications
- **Image forgery localization and detection**:
  - Development of methods for detecting tampered images and localizing the manipulated regions

Which image is manipulated ?

Research field of interest

- **Multimedia Forensics**:
  - Analysis of multimedia data for forensic applications
- **Image forgery localization and detection**:
  - Development of methods for detecting tampered images and localizing the manipulated regions

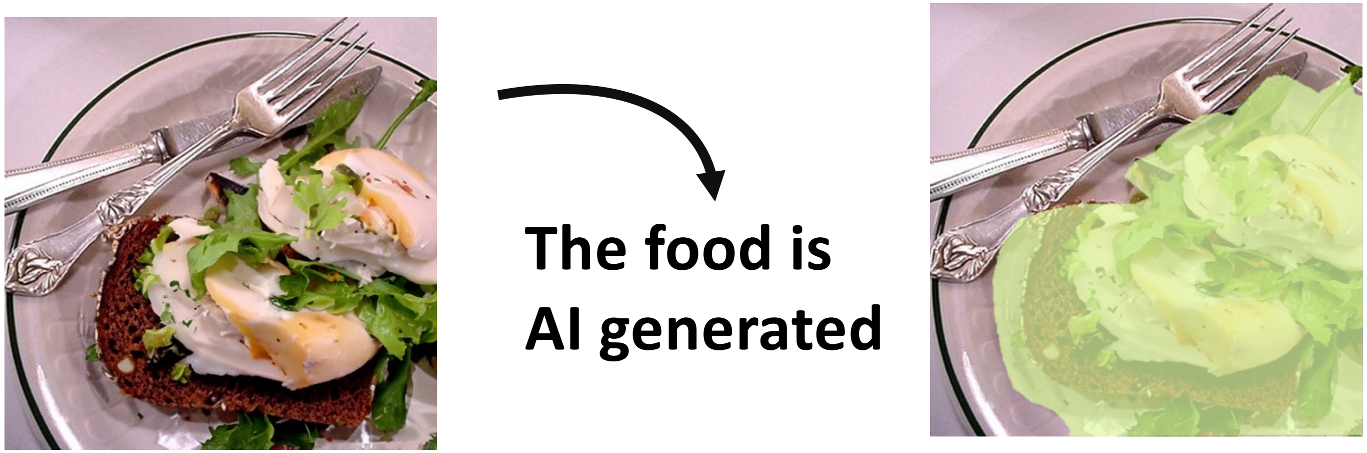
Which image is manipulated ?



Research field of interest

- **Multimedia Forensics**:
  - Analysis of multimedia data for forensic applications
- **Image forgery localization and detection**:
  - Development of methods for detecting tampered images and localizing the manipulated regions

Why?



The food is AI generated

Research field of interest

- **Multimedia Forensics**:
  – Analysis of multimedia data for forensic applications
- **Synthetic Image Detection**:
  – Development of methods for detecting fully synthetic images

Which image is synthetic ?

Research field of interest

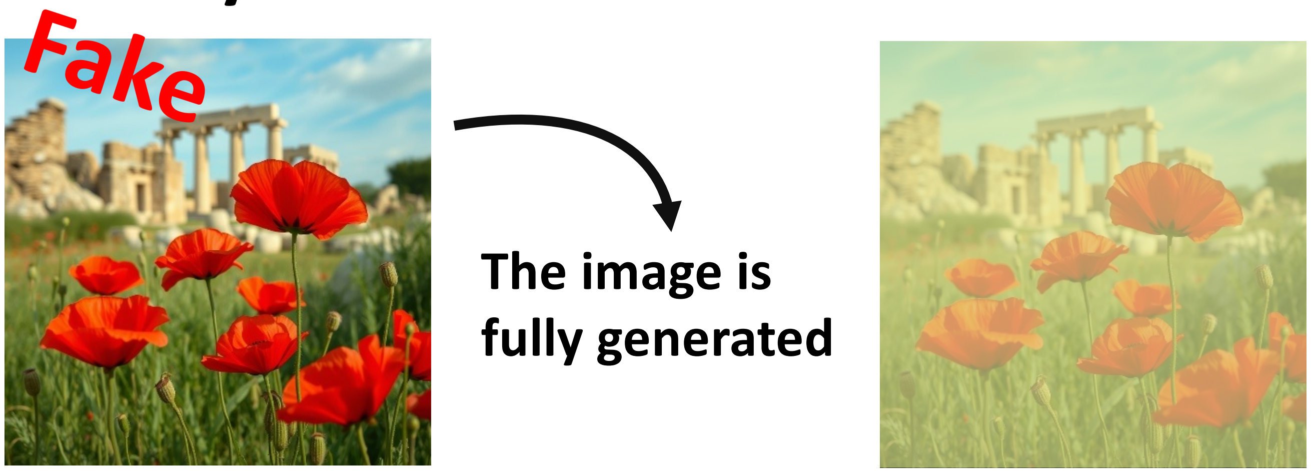
- **Multimedia Forensics**:
  - Analysis of multimedia data for forensic applications
- **Synthetic Image Detection**:
  - Development of methods for detecting fully synthetic images

Which image is synthetic ?

Research field of interest

- **Multimedia Forensics**:
  - Analysis of multimedia data for forensic applications
- **Synthetic Image Detection**:
  - Development of methods for detecting fully synthetic images

Why?



The image is fully generated

Research results

Medical and scientific images:

- Creation of **M3Dsynth**, a medical dataset with AI-generated manipulations, that provides a comprehensive benchmark for image **localization** and **detection** methods
- Design of a **Fusion Method** that produces an accurate pixel-level probability map for scientific image forgery localization

Natural images:

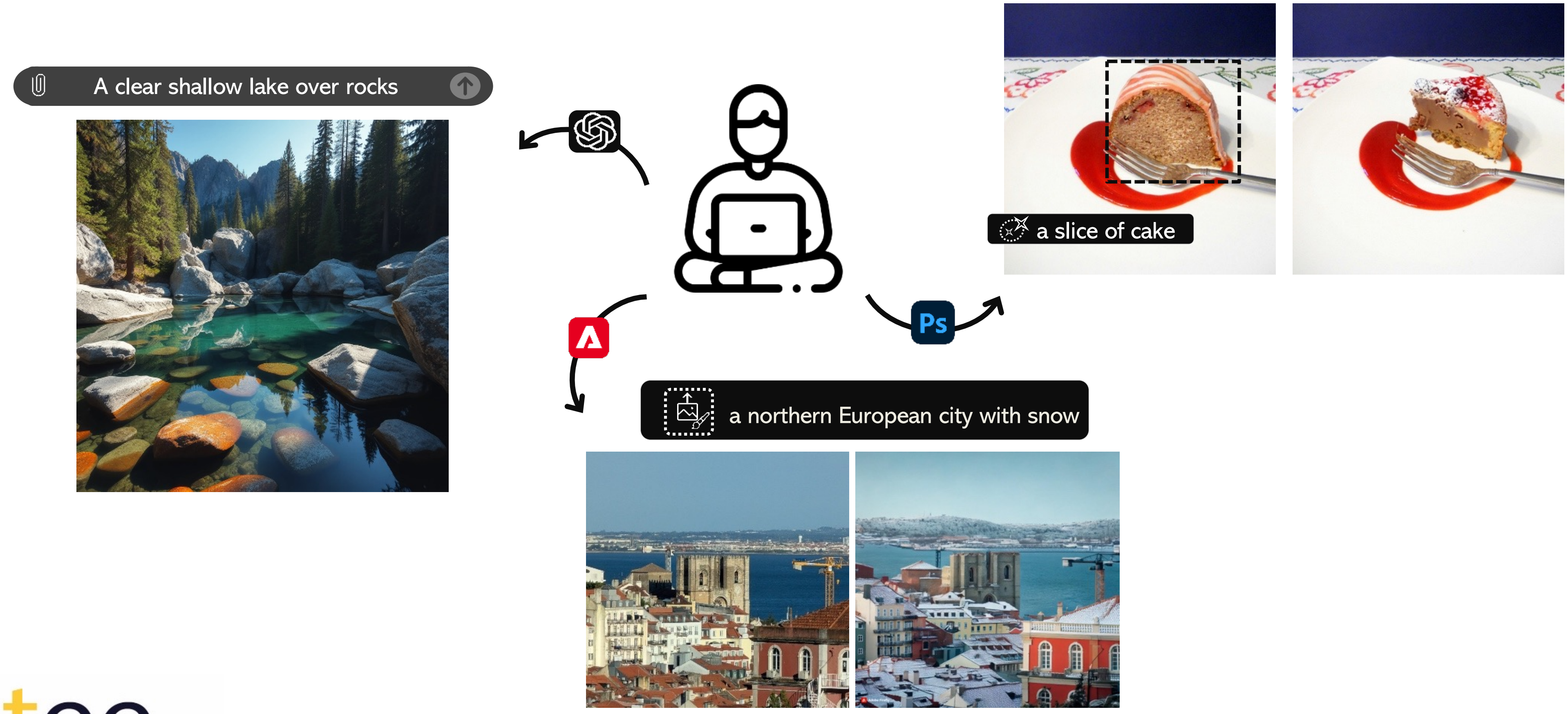
- Analysis of SoTA methods for **diffusion-generated** image detection
- Development of a general and robust detection strategy (**B-Free**) with a new training paradigm and a novel augmentation

Research products

[P1]	R.Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, On the detection of synthetic images generated by diffusion models, **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, Rhodes, Greece, June 2023, pp. 1-5, Publisher, DOI: 10.1109/ICASSP49357.2023.10095167.
[P2]	G. Zingarini, D. Cozzolino, R. Corvi, G. Poggi, and L. Verdoliva, M3Dsynth: A dataset of medical 3D images with AI-generated local manipulations, **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),** Seoul, South Korea, Apr. 2024, pp. 13176-13180, Publisher, DOI: 10.1109/ICASSP48485.2024.10446605.
[P3]	F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A bias-free training paradigm for more general ai-generated image detection, **IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR),** Nashville, Tennessee, USA, June 2025, pp. 18685-18694, Publisher, DOI: 10.1109/CVPR52734.2025.01741.

PhD thesis: Overview

- Problem
  - AI generative tools are now easily accessible by any user with internet connection


A clear shallow lake over rocks
a slice of cake
Ps
a northern European city with snow

PhD thesis: Overview

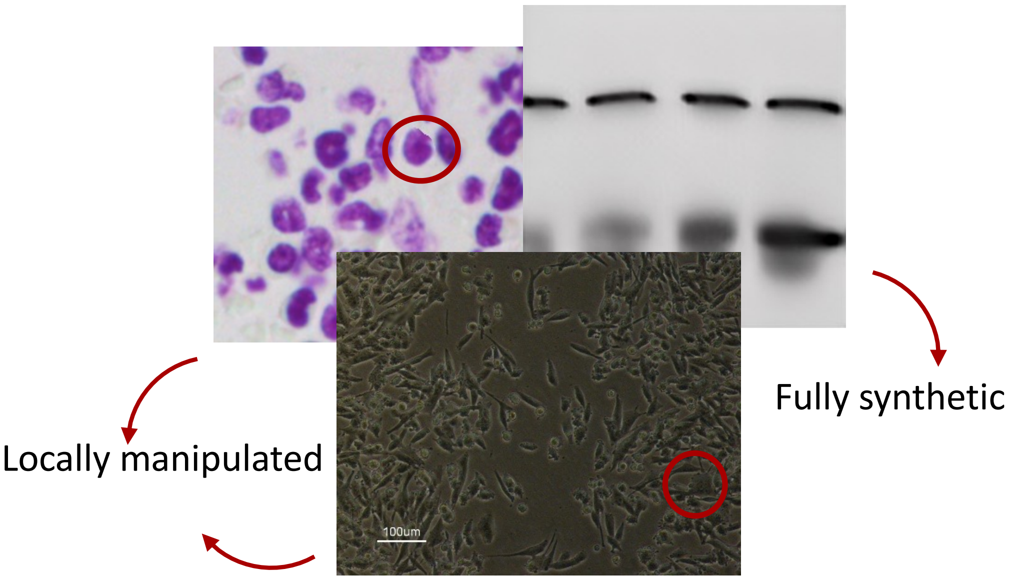
- **Synthetic image generation** concerns many fields:

Natural images

Fully synthetic

Locally manipulated

Medical/Scientific images

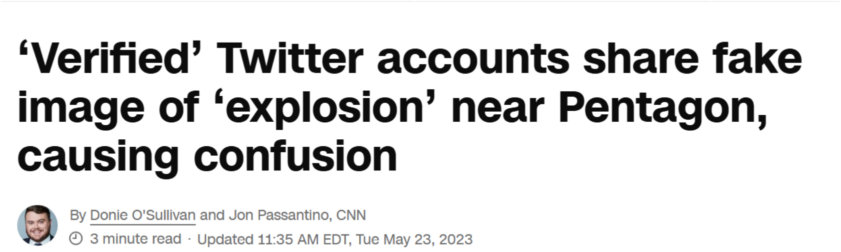


Locally manipulated

Fully synthetic

PhD thesis: Overview

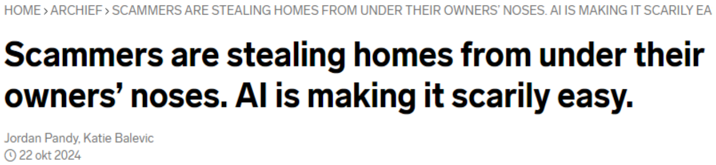
- Problem
  - AI tools can be maliciously used to spread **disinformation**



‘Verified’ Twitter accounts share fake image of ‘explosion’ near Pentagon, causing confusion

By Donie O'Sullivan and Jon Passantino, CNN

3 minute read · Updated 11:35 AM EDT, Tue May 23, 2023

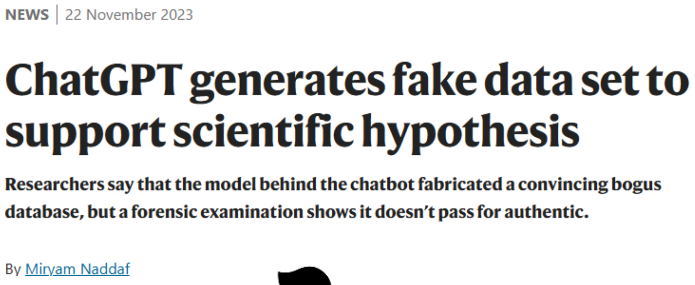


HOME › ARCHIEF › SCAMMERS ARE STEALING HOMES FROM UNDER THEIR OWNERS' NOSES. AI IS MAKING IT SCARILY EA

Scammers are stealing homes from under their owners’ noses. AI is making it scarily easy.

Jordan Pandy, Katie Balevic

22 okt 2024

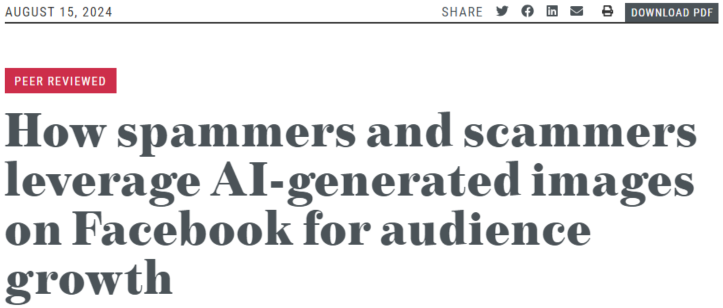


NEWS | 22 November 2023

ChatGPT generates fake data set to support scientific hypothesis

Researchers say that the model behind the chatbot fabricated a convincing bogus database, but a forensic examination shows it doesn’t pass for authentic.

By Miryam Naddaf



AUGUST 15, 2024

SHARE

DOWNLOAD PDF

PEER REVIEWED

How spammers and scammers leverage AI-generated images on Facebook for audience growth

PhD thesis: Overview

- Problem
  - Malicious use of Generative AI
  - Spread of **disinformation** and **alterations** of **scientific results**

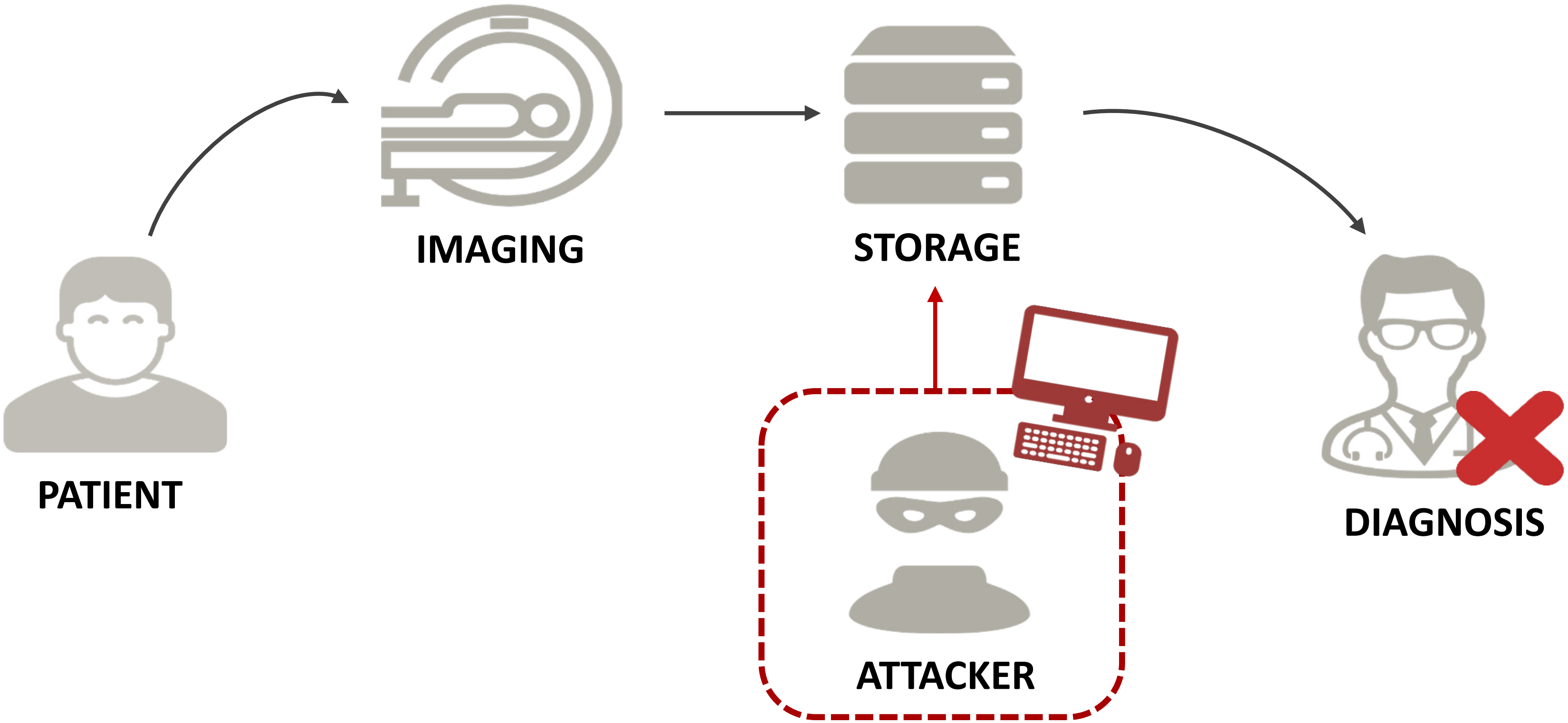
- Objective

  How to face the problem?

  - Generation of medical data to train learning-based **image forgery localization** models **(M3Dsynth)**
  - Design of a **detector for natural synthetic images** robust to the sharing process over social networks **(B-Free)**

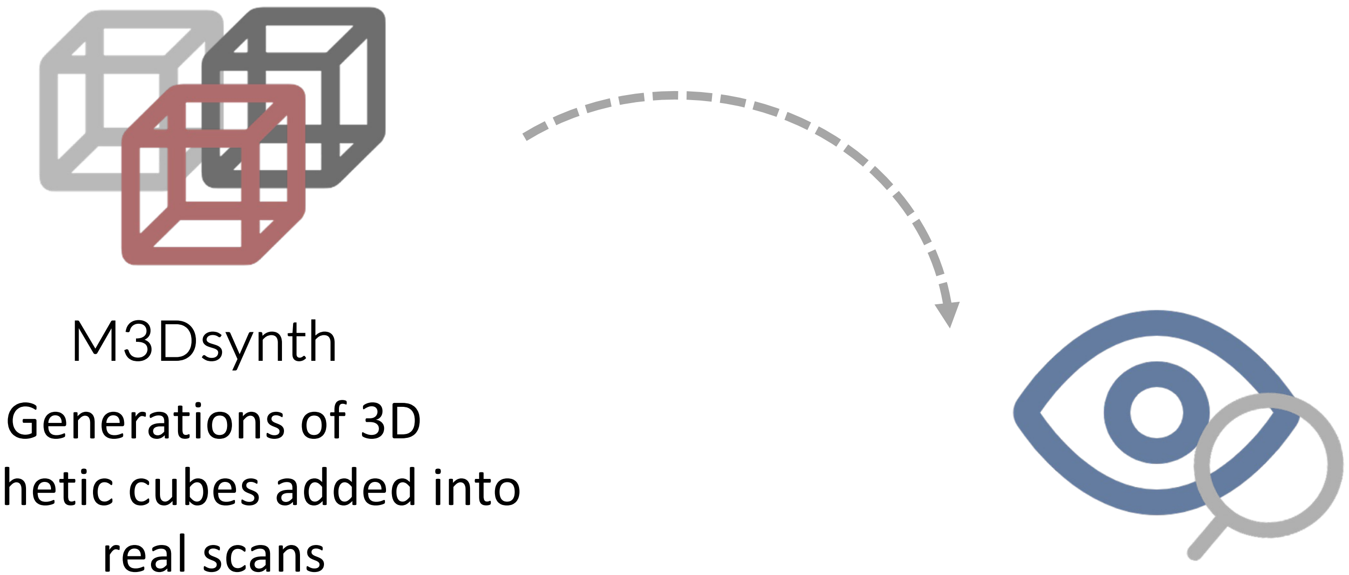
Forgery Localization: M3Dsynth

- An **attacker** enters the PACS and induces an incorrect diagnosis
- **AI-Generative models can be used** to manipulate such images



M3Dsynth: Overview

- Generation of the **M3Dsynth** dataset of 3D CT lung scans with local manipulations
- **Benchmark** analysis of the SoTA detection and localization methods



M3Dsynth

Generations of 3D synthetic cubes added into real scans

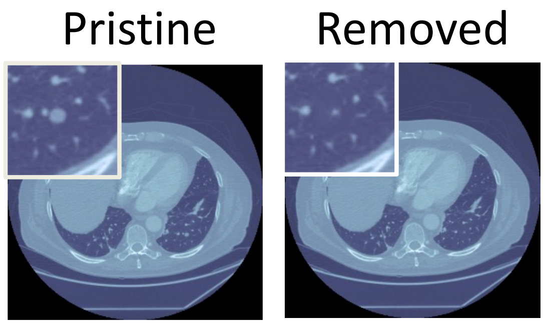
Benchmark

Testing models for medical manipulations

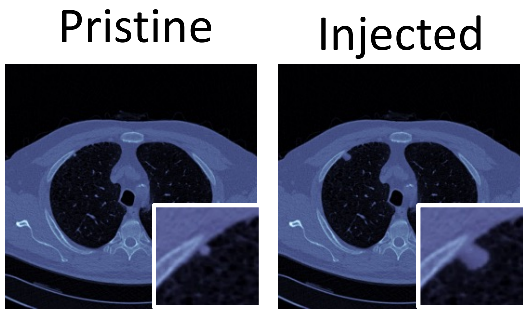
M3Dsynth: Methodology

- **M3Dsynth** consists of 8,577 manipulated samples with **injection** or **removal** of a cancer nodule

Removal Task: the real malignant nodule is replaced with a fake benign nodule with a diameter less than 8 mm

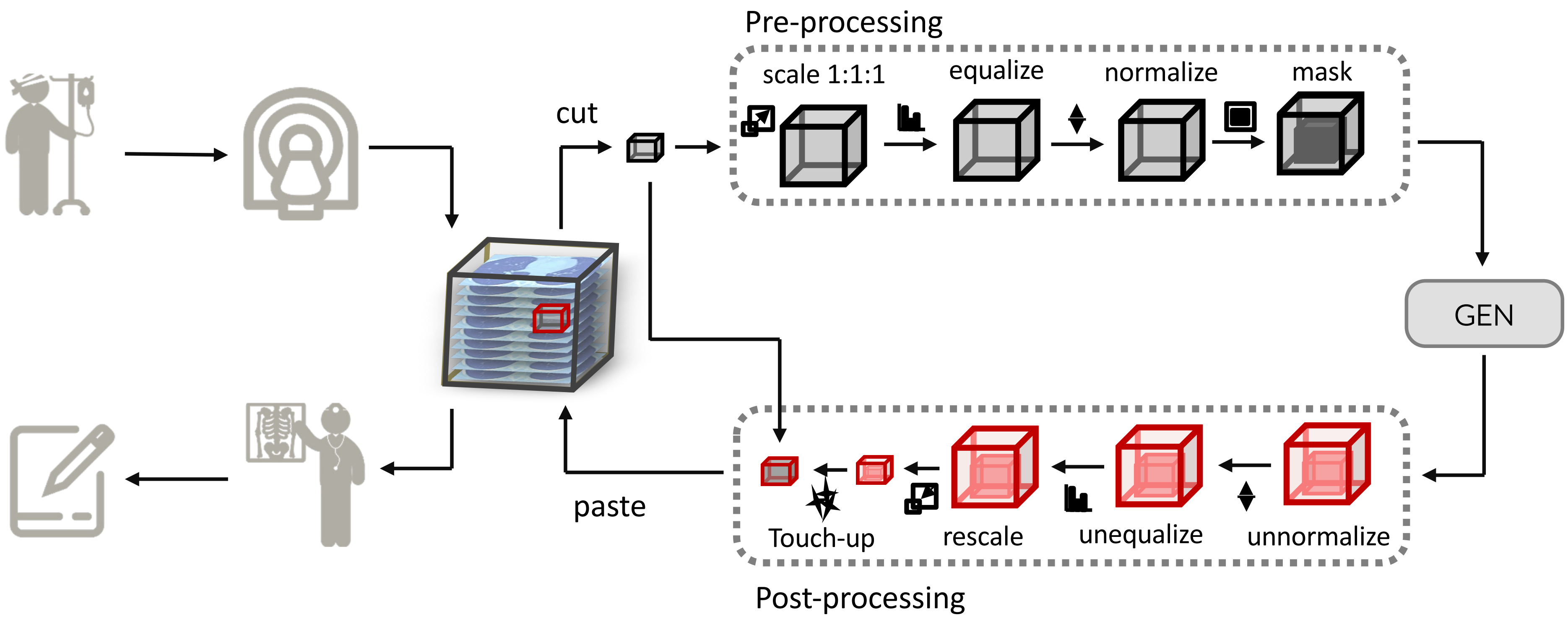


Injection Task: a fake malignant nodule with a diameter over than 10 mm is generated



M3Dsynth: Methodology

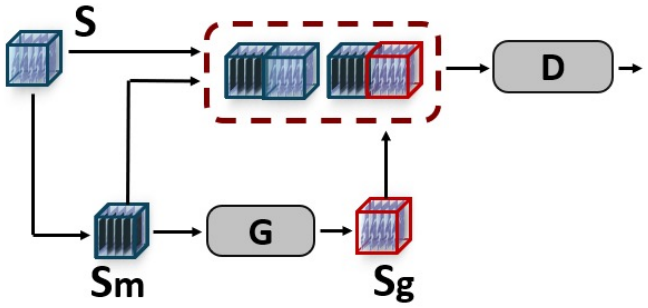
- The tampering process works on 32-mm cubes selected from the original CT-scan at the desired location



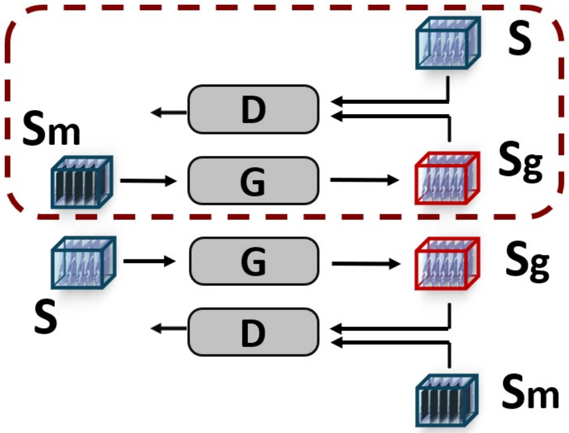
M3Dsynth: Methodology

- **Three** versions of the same manipulated CT scan using different AI-generative model methods

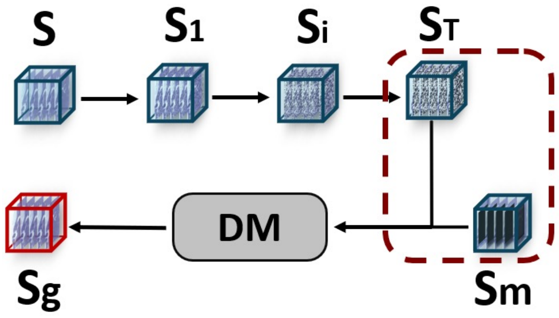
Pix2Pix GAN



Cycle GAN



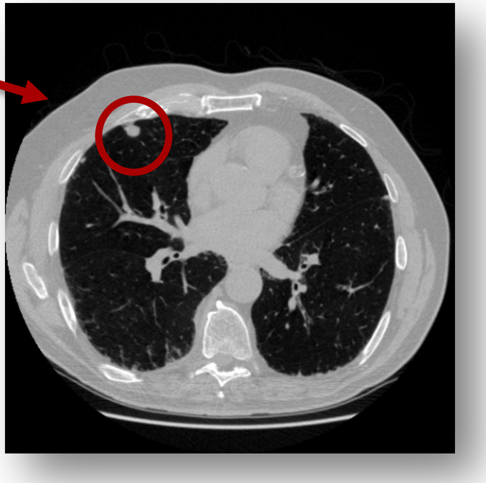
Diffusion Model



M3Dsynth: Qualitative Analysis

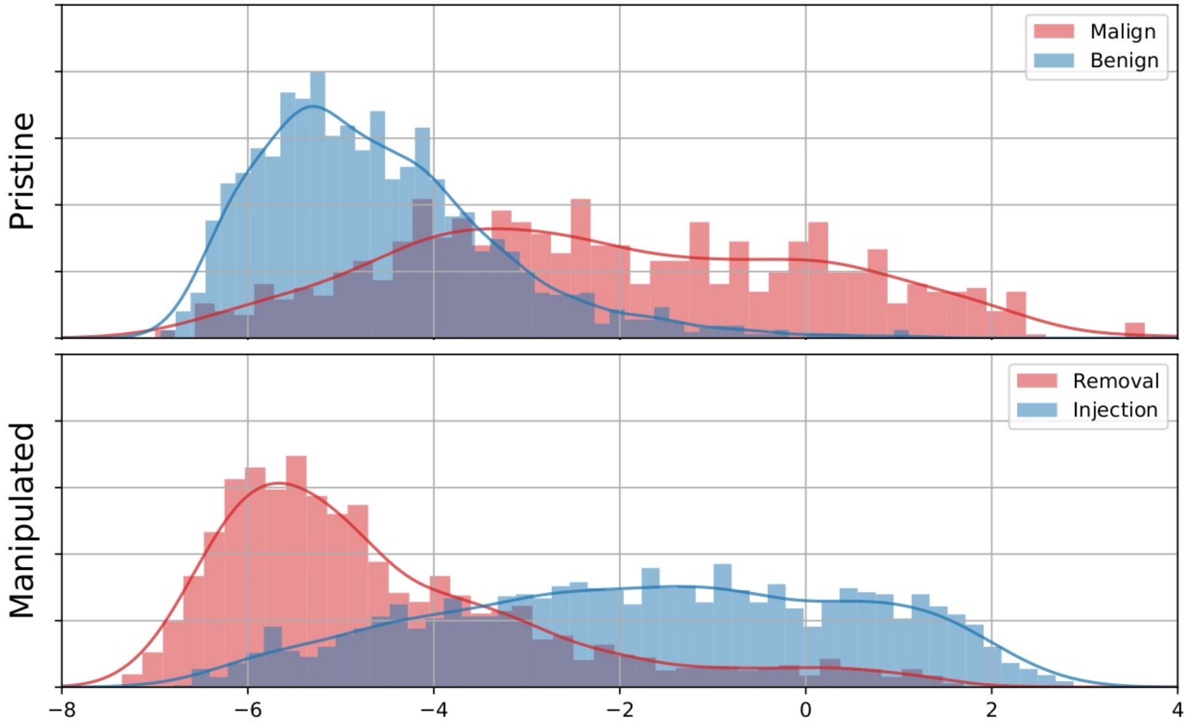
- Evaluation of the generated images through a **computer-aided diagnostic tool**
- Histograms of pristine and manipulated scans provide **inverted diagnosis**

INJECTED NODULE
MALIGNACY SCORE:
0.80



M3Dsynth: Qualitative Analysis

- Evaluation of the generated images through a **computer-aided diagnostic tool**
- Histograms of pristine and manipulated scans provide **inverted diagnoses**



M3Dsynth: Benchmark

Methods using **compression artifacts** or internal **camera processing traces** are not considered

Method	RGB Img	Others	Task
Xception	**Y**	**None**	**Det.**
U-Net	**Y**	**None**	**Det. & Loc.**
HP-FCN	**N**	**HP filters**	**Det. & Loc.**
ManTraNet	**Y**	**HP filters**	**Det. & Loc.**
MVSS-Net	**Y**	**Trainable HP filter**	**Det. & Loc.**
TruFor	**Y**	**Noiseprint++**	**Det. & Loc.**

M3Dsynth: Results

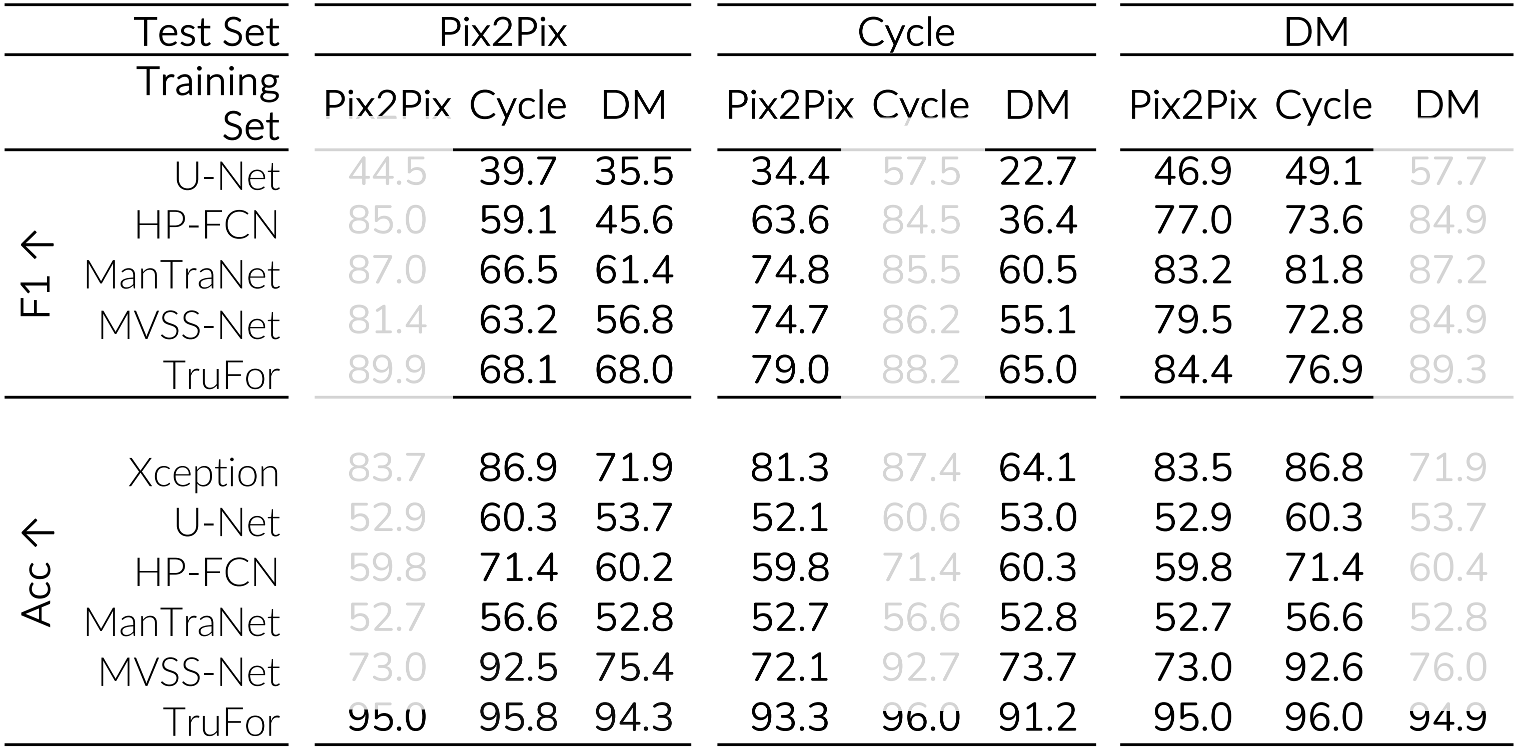
- **Localization**: the performance is good on average (especially Trufor, ManTraNet)
- **Detection**: several methods have good detection performance showing lower results only in a few cases (HP-FCN and U-Net)

	Test Set	Pix2Pix			Cycle			DM		
	Training Set	Pix2Pix	Cycle	DM	Pix2Pix	Cycle	DM	Pix2Pix	Cycle	DM
F1 ↑	U-Net	44.5	39.7	35.5	34.4	57.5	22.7	46.9	49.1	57.7
	HP-FCN	85.0	59.1	45.6	63.6	84.5	36.4	77.0	73.6	84.9
	ManTraNet	87.0	66.5	61.4	74.8	85.5	60.5	83.2	81.8	87.2
	MVSS-Net	81.4	63.2	56.8	74.7	86.2	55.1	79.5	72.8	84.9
	TruFor	89.9	68.1	68.0	79.0	88.2	65.0	84.4	76.9	89.3
Acc ↑	Xception	83.7	86.9	71.9	81.3	87.4	64.1	83.5	86.8	71.9
	U-Net	52.9	60.3	53.7	52.1	60.6	53.0	52.9	60.3	53.7
	HP-FCN	59.8	71.4	60.2	59.8	71.4	60.3	59.8	71.4	60.4
	ManTraNet	52.7	56.6	52.8	52.7	56.6	52.8	52.7	56.6	52.8
	MVSS-Net	73.0	92.5	75.4	72.1	92.7	73.7	73.0	92.6	76.0
	TruFor	95.0	95.8	94.3	93.3	96.0	91.2	95.0	96.0	94.9

Aligned Data Results

M3Dsynth: Results

- We test the **generalization ability** by testing each generator against all the others
- Only a limited impairment is observed on a **non-aligned scenario**

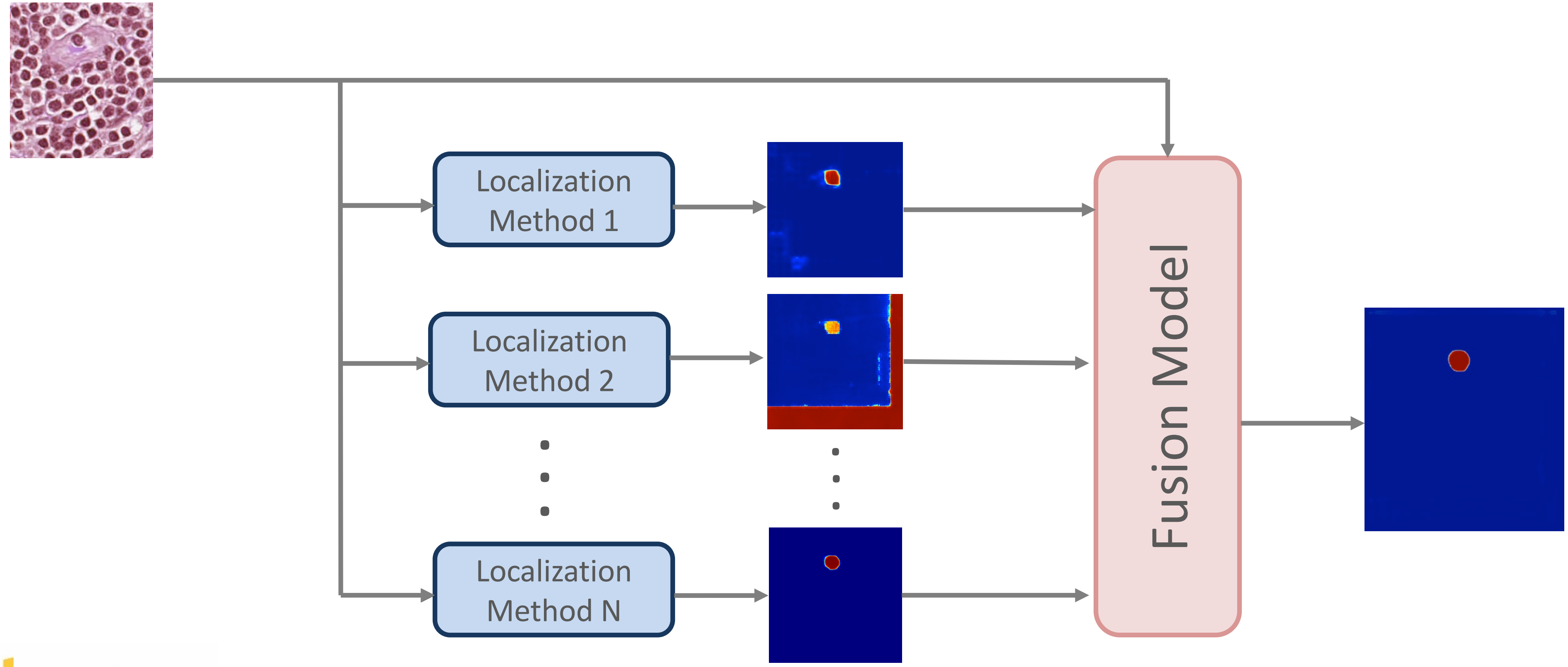


	Test Set	Pix2Pix			Cycle			DM		
	Training Set	Pix2Pix	Cycle	DM	Pix2Pix	Cycle	DM	Pix2Pix	Cycle	DM
F1 ↑	U-Net	44.5	39.7	35.5	34.4	57.5	22.7	46.9	49.1	57.7
	HP-FCN	85.0	59.1	45.6	63.6	84.5	36.4	77.0	73.6	84.9
	ManTraNet	87.0	66.5	61.4	74.8	85.5	60.5	83.2	81.8	87.2
	MVSS-Net	81.4	63.2	56.8	74.7	86.2	55.1	79.5	72.8	84.9
	TruFor	89.9	68.1	68.0	79.0	88.2	65.0	84.4	76.9	89.3
Acc ↑	Xception	83.7	86.9	71.9	81.3	87.4	64.1	83.5	86.8	71.9
	U-Net	52.9	60.3	53.7	52.1	60.6	53.0	52.9	60.3	53.7
	HP-FCN	59.8	71.4	60.2	59.8	71.4	60.3	59.8	71.4	60.4
	ManTraNet	52.7	56.6	52.8	52.7	56.6	52.8	52.7	56.6	52.8
	MVSS-Net	73.0	92.5	75.4	72.1	92.7	73.7	73.0	92.6	76.0
	TruFor	95.0	95.8	94.3	93.3	96.0	91.2	95.0	96.0	94.9

Cross Validation Scenario

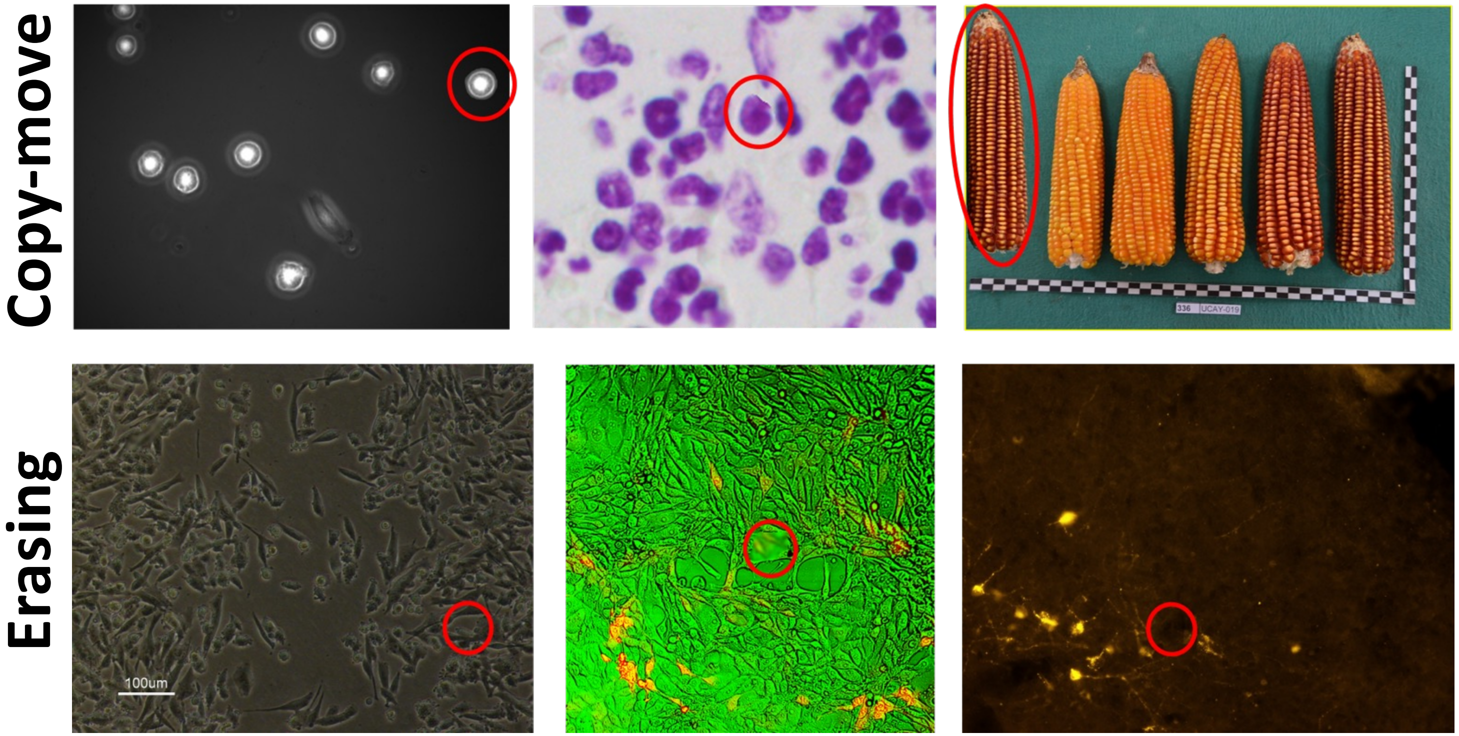
Forgery Localization: Fusion Method

- **Fusion Method** for Scientific Forgery Localization Task
- A final **improved probability map** that identify the manipulation



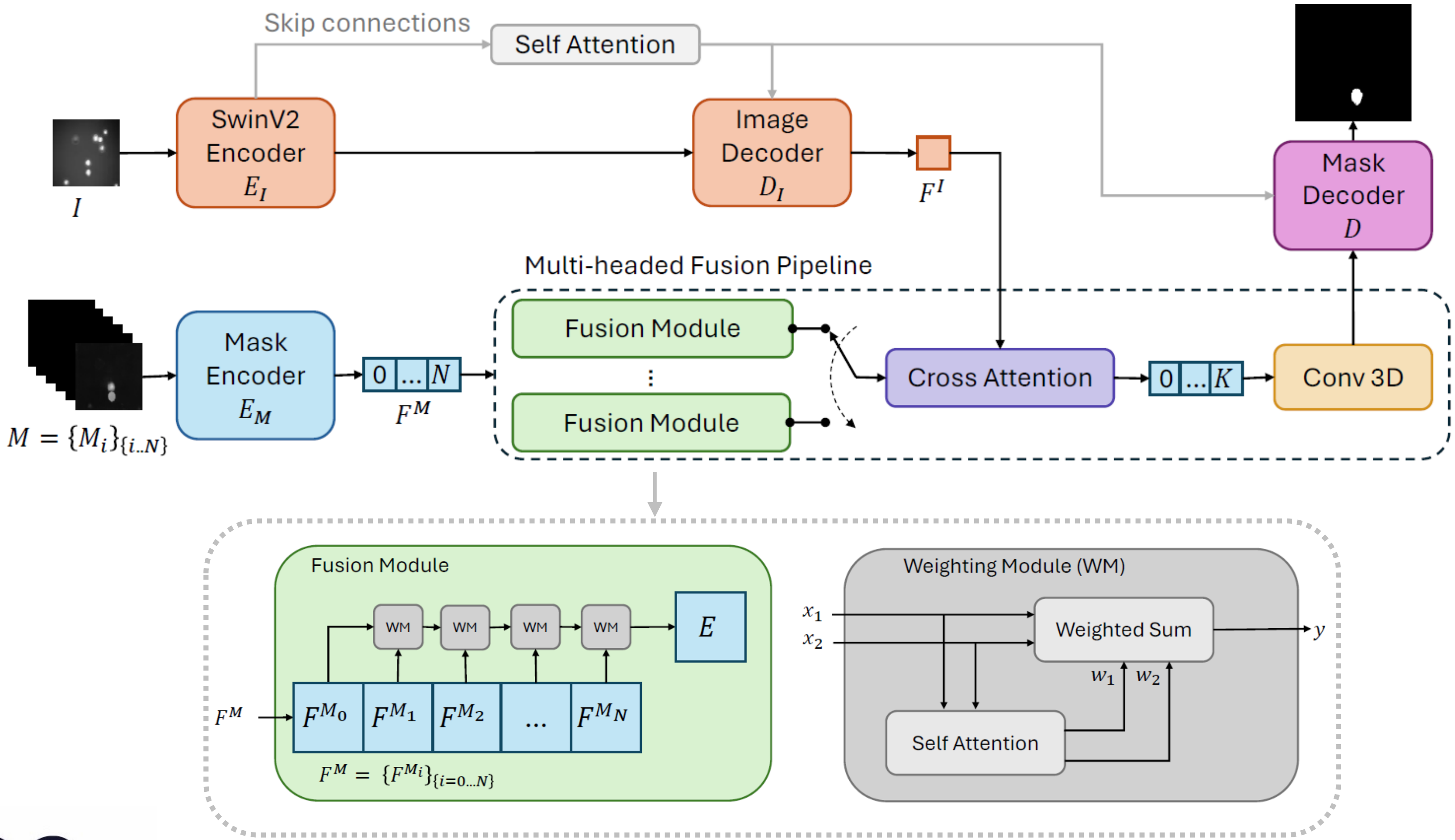
Fusion Method: Methodology

- Dataset of 20K images with copy-move and inpainting
- Highly **heterogeneous** with a wide variety of laboratory sources



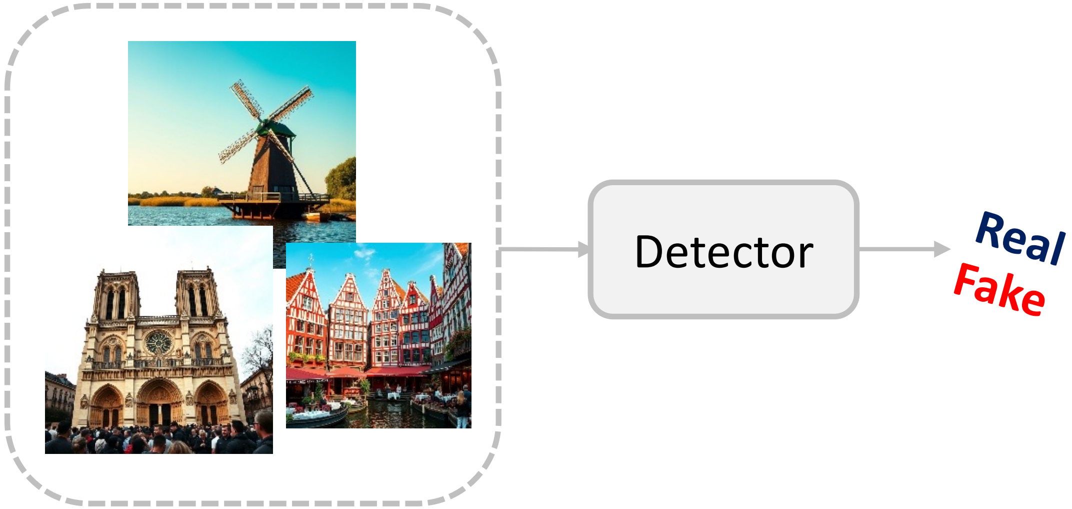
Fusion Method: Methodology

- The **Fusion** Architecture takes in input a variable number of maps



Synthetic Detection: B-Free

Synthetic Image Detection for natural images

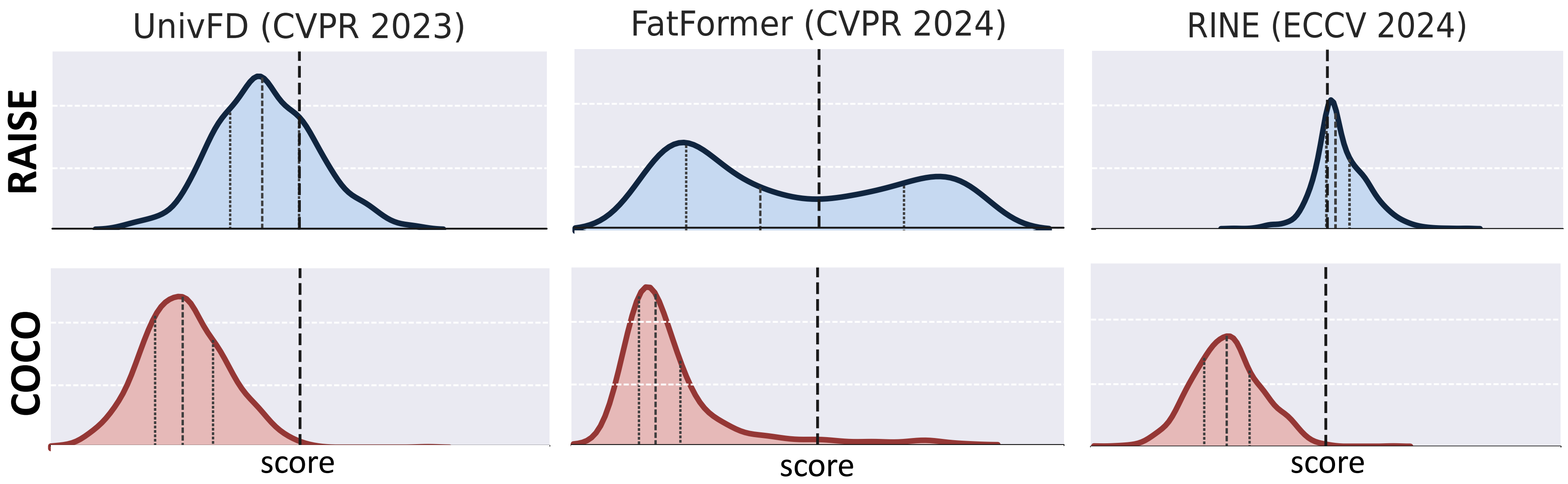


- Training datasets can be affected by **biases**
- Impacting the detector’s ability to **generalize**

Synthetic Detection: B-Free

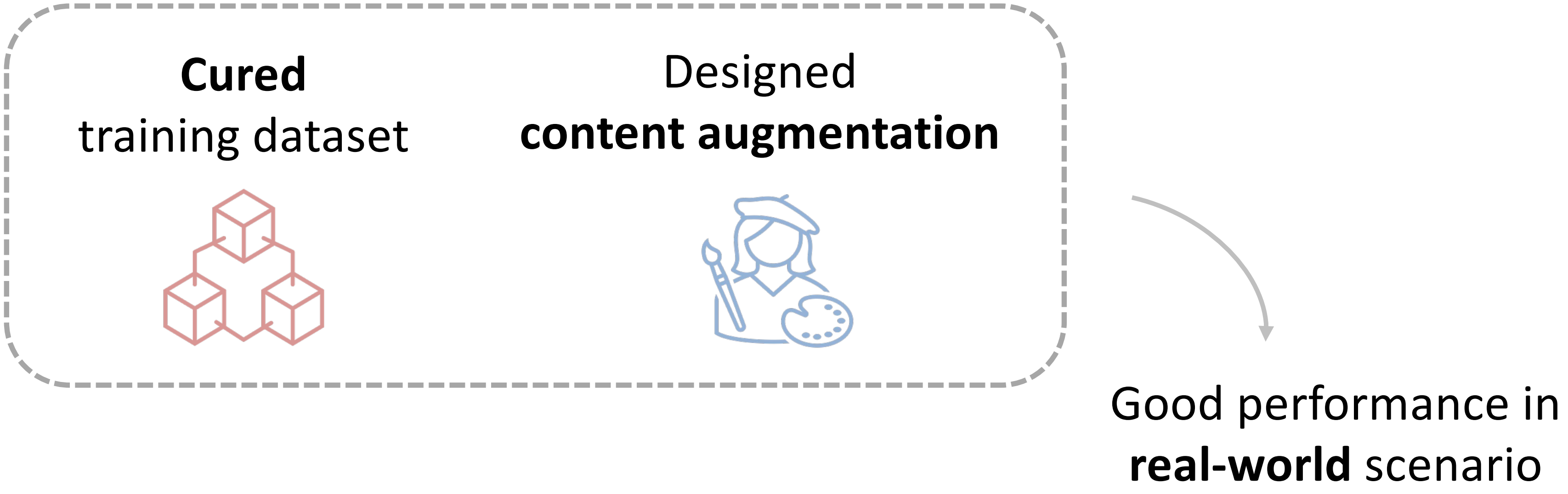
- Training datasets can be affected by **biases**
- Impacting the detector’s ability to **generalize**

DALL·E 3



B-Free: Overview

- A good forensic detector should detect generative artifacts rather than reflect data biases
- We propose **B-Free**: a new training paradigm to prevent biases



Cured training dataset

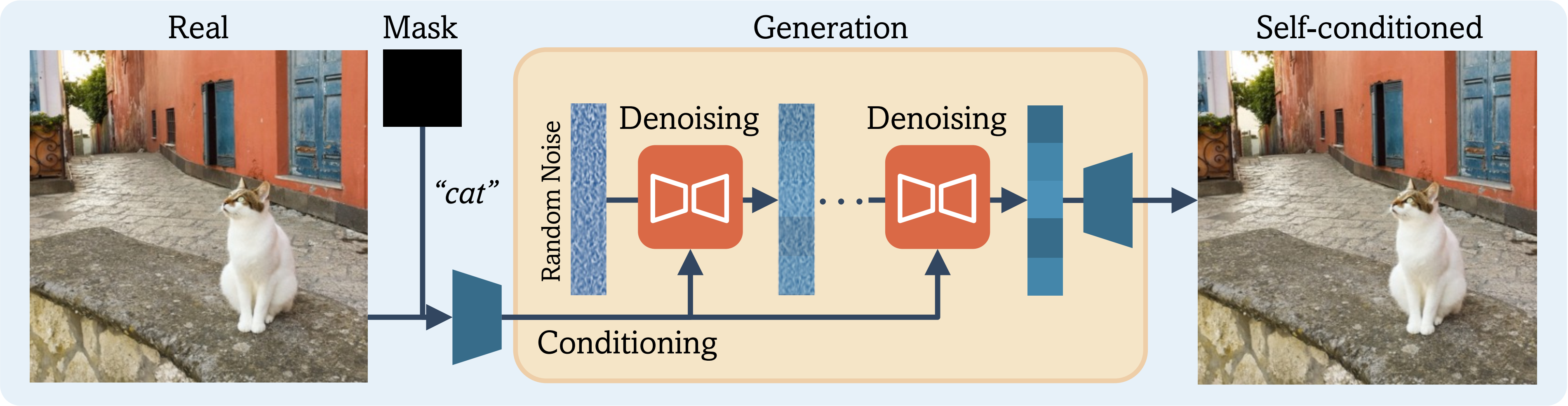
Designed **content augmentation**

Good performance in **real-world** scenario

B-Free: Methodology

- Our new training paradigm relies on the use of:

🔧 **Self-conditioned reconstructions** of real images using SD 2.1

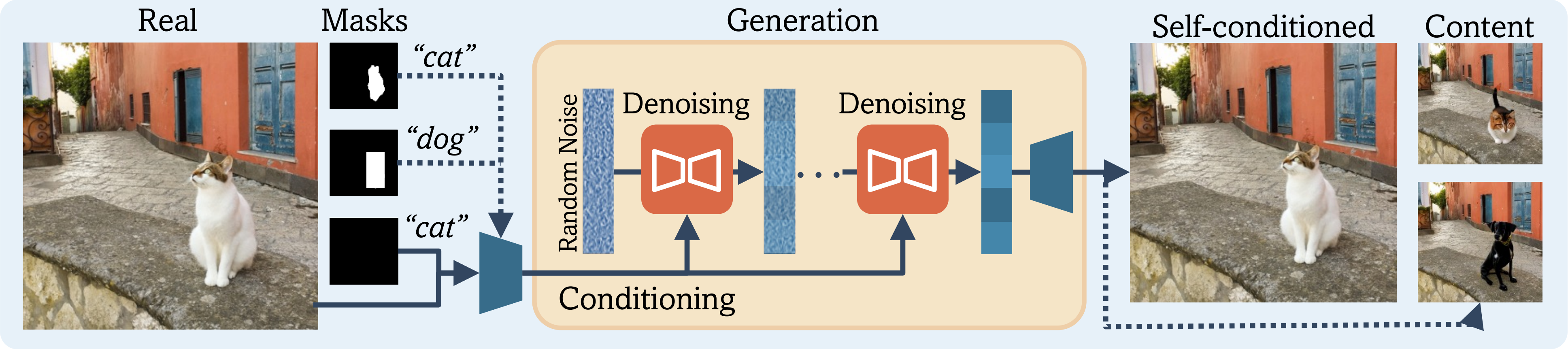


B-Free: Methodology

- Our new training paradigm relies on the use of:

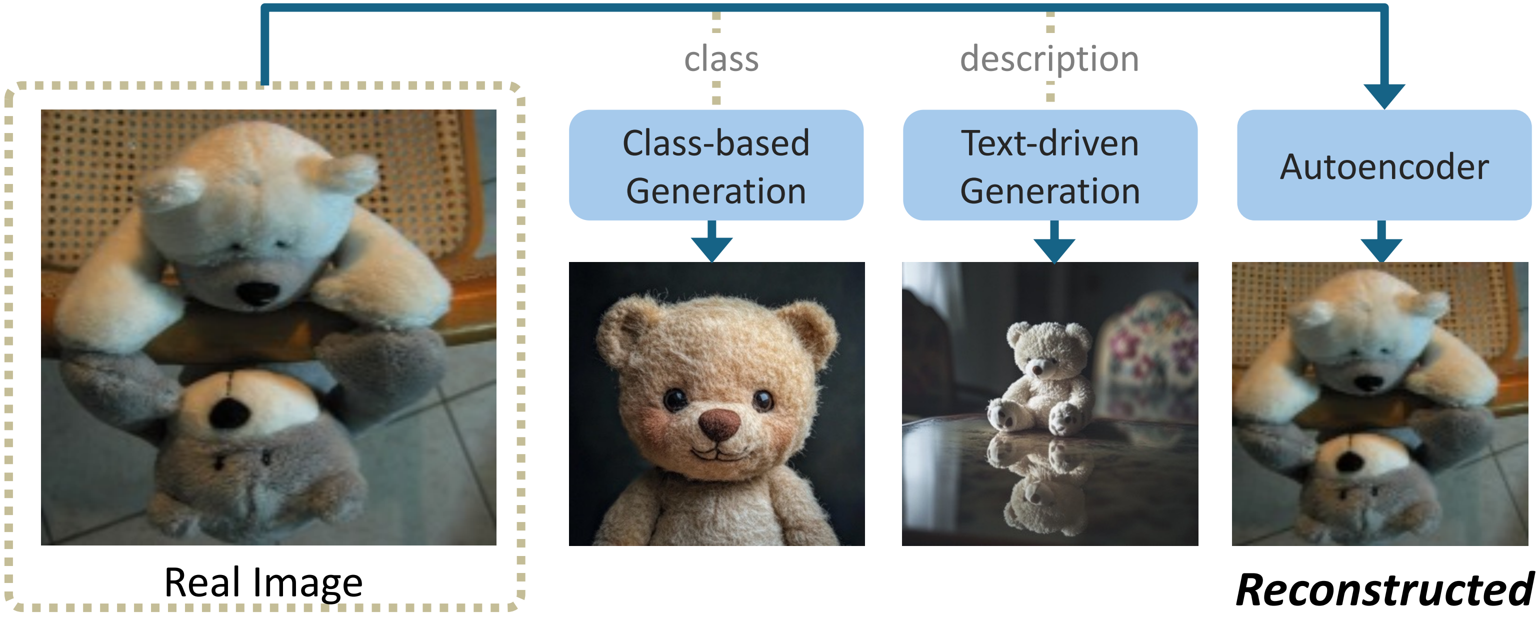
🔧 **Self-conditioned reconstructions** of real images using SD 2.1

🖌 **Content augmentation:** locally inpainted versions of the images



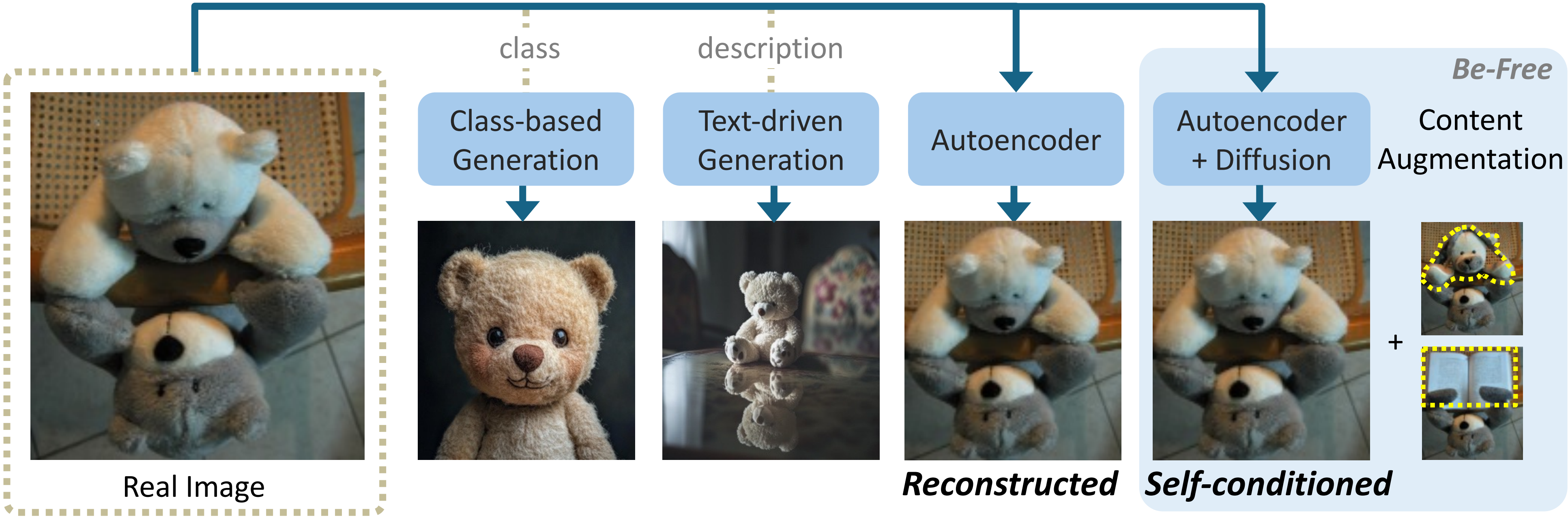
B-Free: Methodology

- The **self-conditioned** image shares the same content as the real one
- **Diffusion steps** are also included to keep the forensic artifacts of the generation process



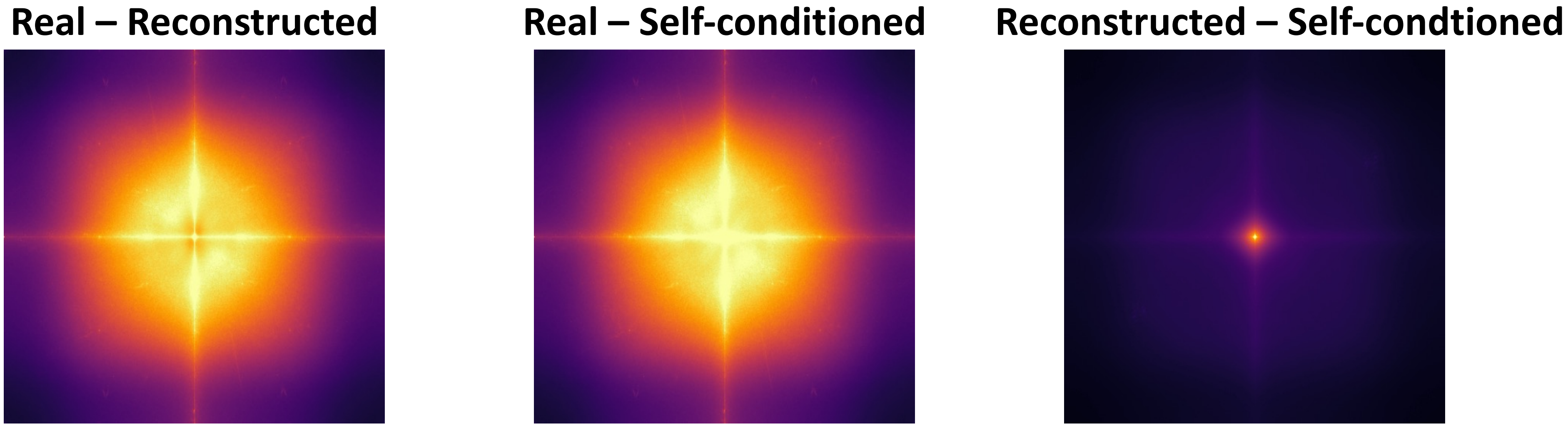
B-Free: Methodology

- The **self-conditioned** image shares the same content as the real one
- **Diffusion steps** are also included to keep the forensic artifacts of the generation process



B-Free: Overview

- The use of diffusion steps allows to exploit the inconsistencies across a **wider range of frequencies**
- Self-conditioned images embed forensics artifacts even at **lower frequencies**



Real – Reconstructed

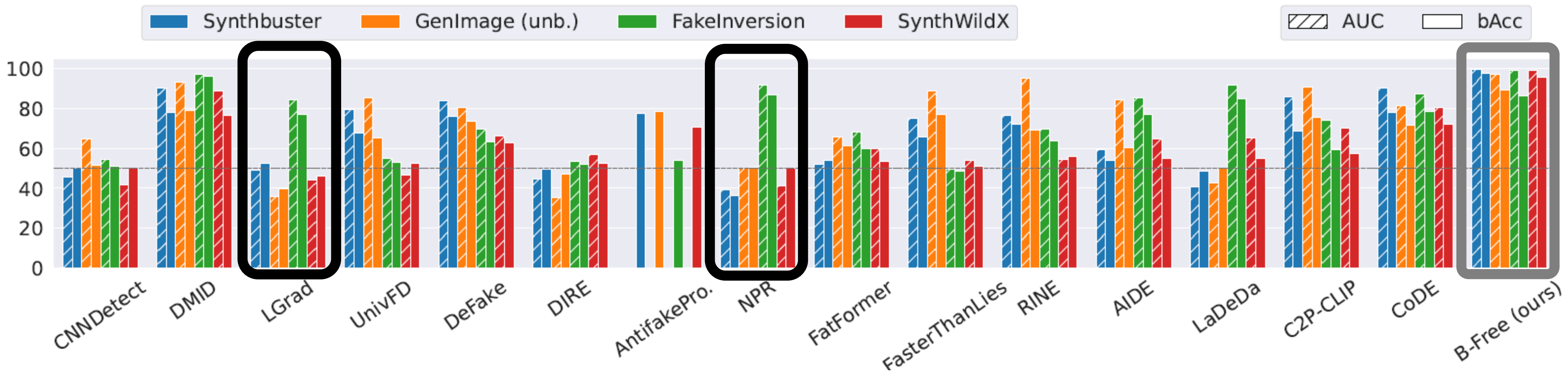
Real – Self-conditioned

Reconstructed – Self-condtioned

Power spectra
(by averaging the **difference** on 2000 images)

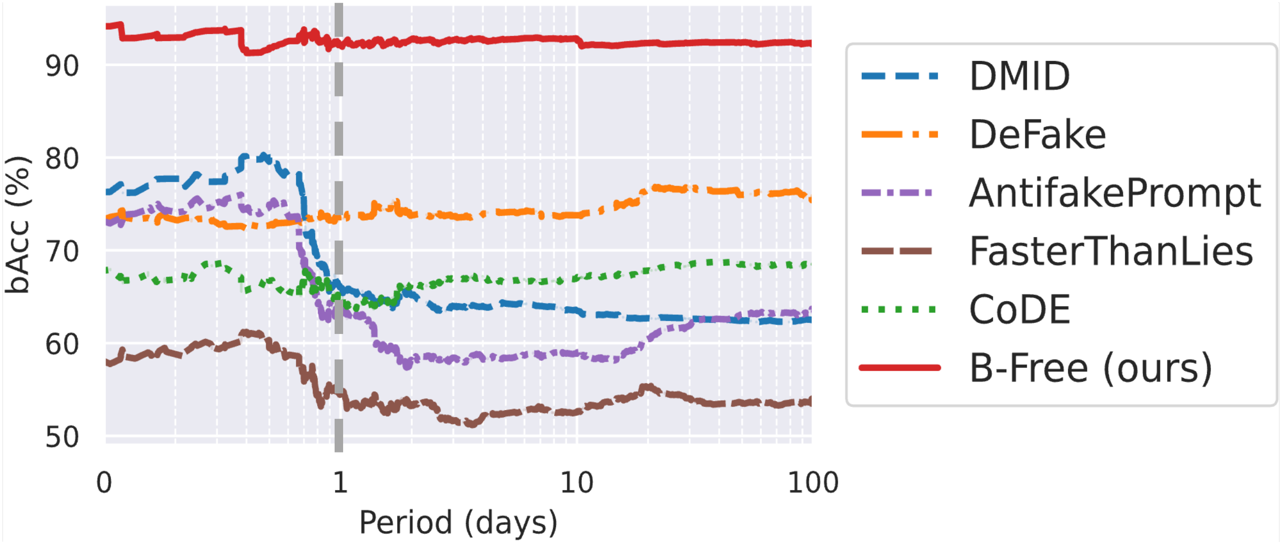
B-Free: Results

- Some methods show good performance on only a single dataset proving the exploitation of data-biases instead of forensic artifacts
- Our method presents **uniform performance** across all datasets



B-Free: Results

- We collected images that went **viral** and evaluated the performance in function of the time elapsed from the first online post
- Their quality is affected after **just one day** of re-posting
- **B-Free** performs always above 92%



Collection of **1400** real/fake images viral over the web

Conclusions

- We explored the **Medical Image Forgery Localization** task, proposing an AI-manipulated dataset to evaluate SoTA models

- We designed a fusion approach that takes as input a variable number of localization maps and generates a more accurate result

- We developed a general and robust **Synthetic Image Detection** method with a bias-free training paradigm

Thank you for the attention!