2024 IEEE International Conference on
Acoustics, Speech and Signal Processing
(ICASSP 2024), Seoul, Korea, 14~19 April

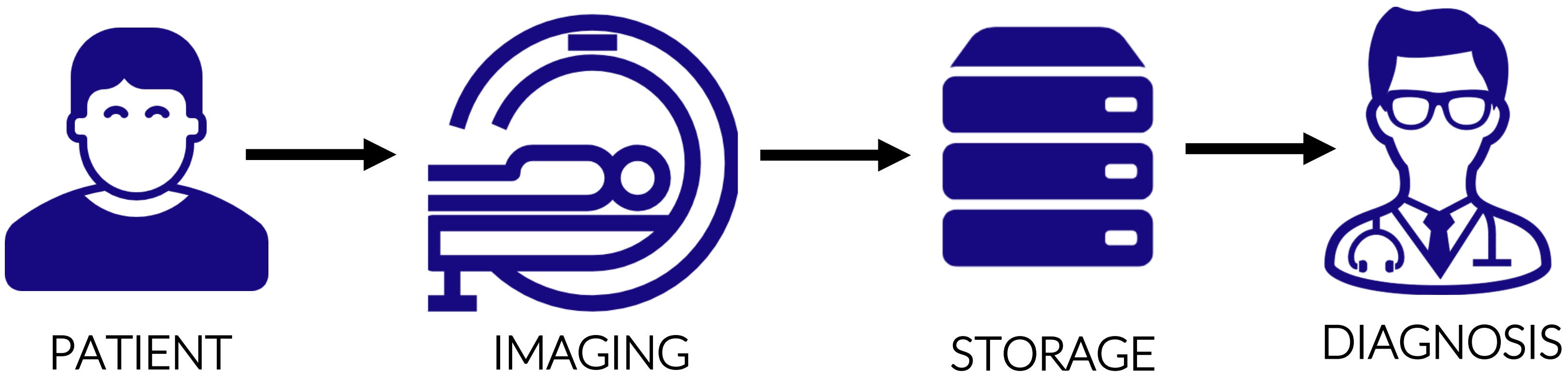
M3Dsynth: A dataset of medical 3D images with AI-generated local manipulations

Authors: Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, Luisa Verdoliva

University Federico II of Naples

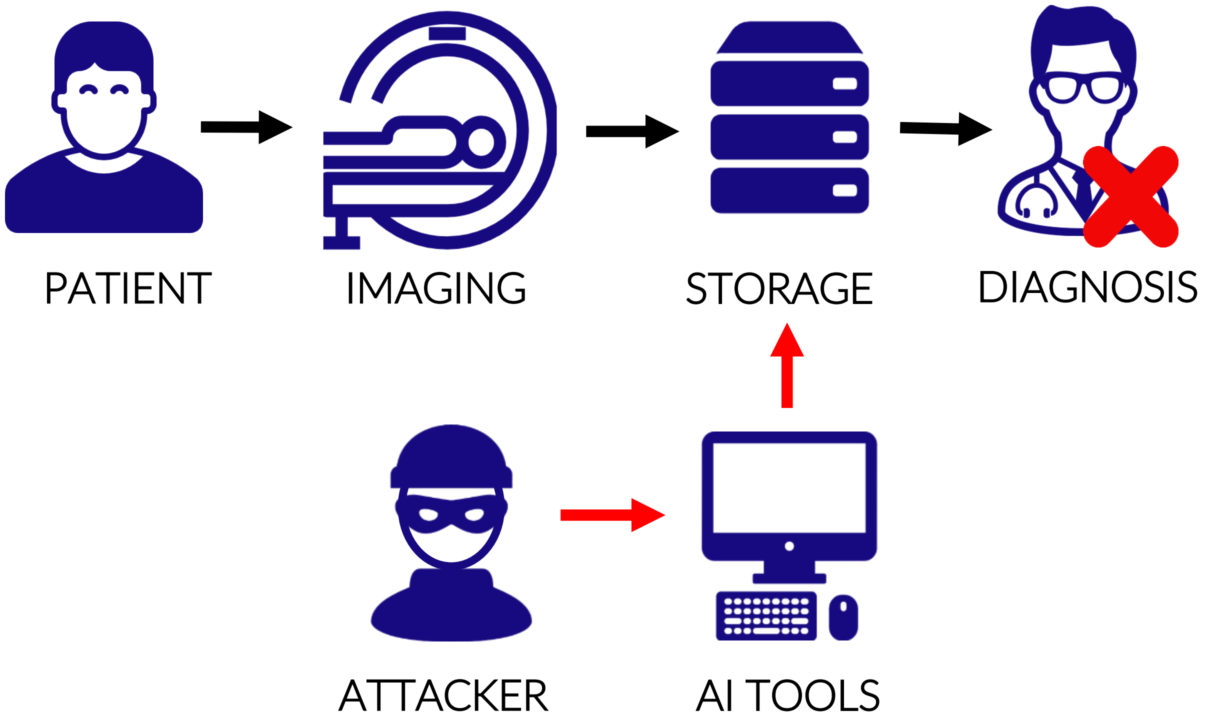
Background

- Most diseases diagnoses rely on medical imaging techniques
- 3D medical images are stored in secure Picture and Archive Communication System (PACS) servers



Background

- An **attacker** could enter the system and modify medical CT scans to induce an incorrect diagnosis [1]



[1] Y. Mirsky et al. “CT-GAN: Malicious tampering of 3d medical imagery using deep learning,” 28th USENIX Security Symposium, 2019

Objective

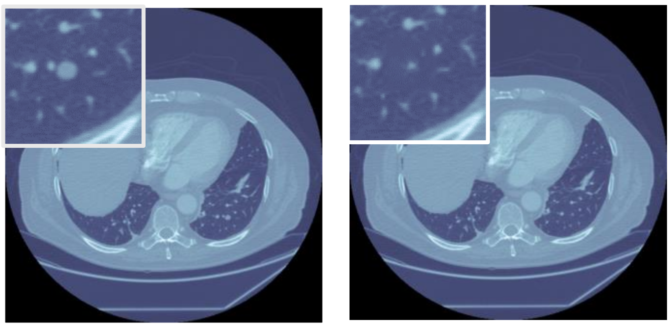
- Most efforts in the **forensics community** are focused on the detection of deepfakes in natural videos/images

- We aim to stimulate the community to pay attention to AI-based manipulations of medical images by proposing **a dataset and a benchmark** [2]

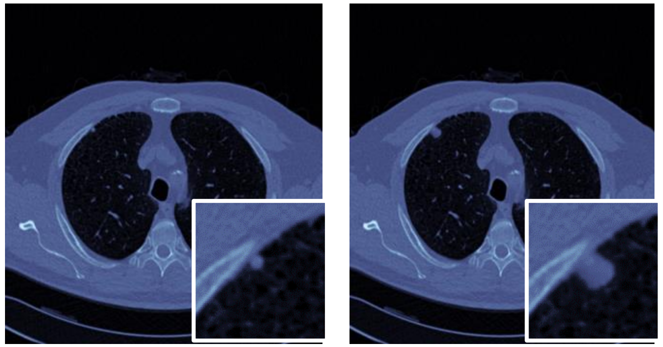
[2] https://grip-unina.github.io/M3Dsynth/

Data generation process

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



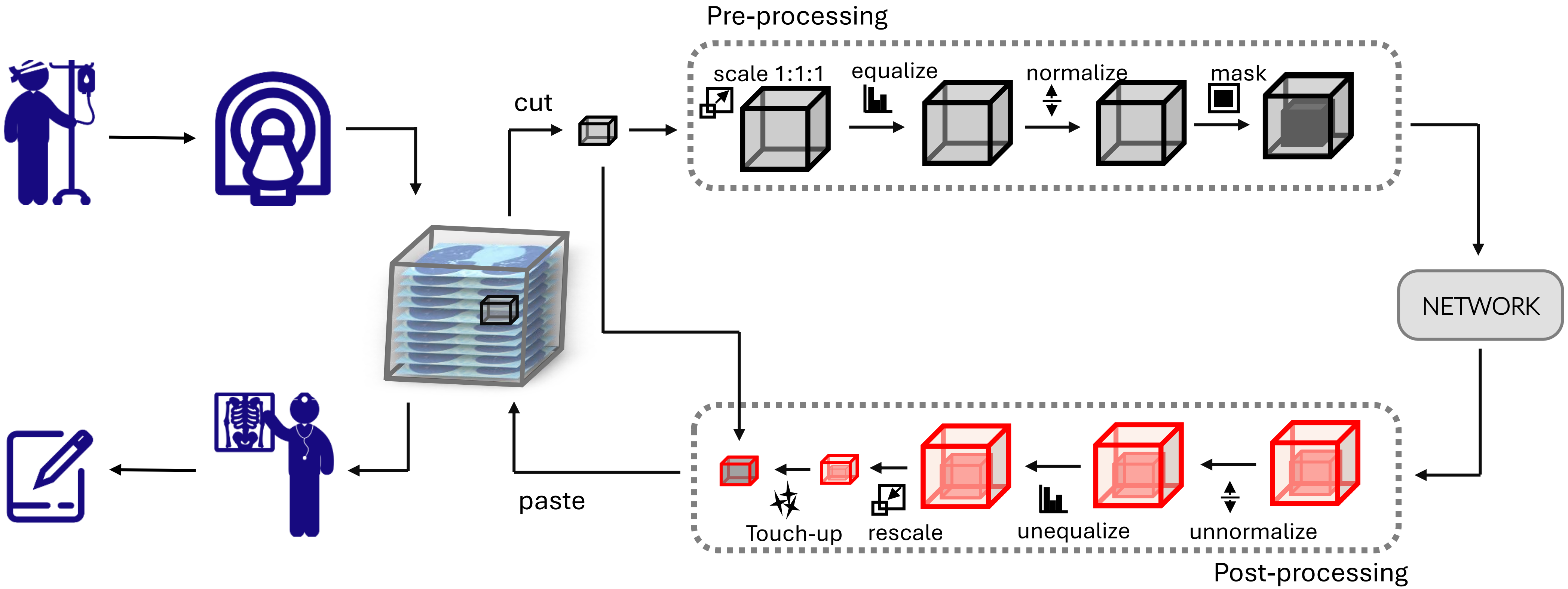
Removal Task: the real malignant nodule is replaced with a fake benign nodule with a diameter less than 8 mm



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

Data generation process

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

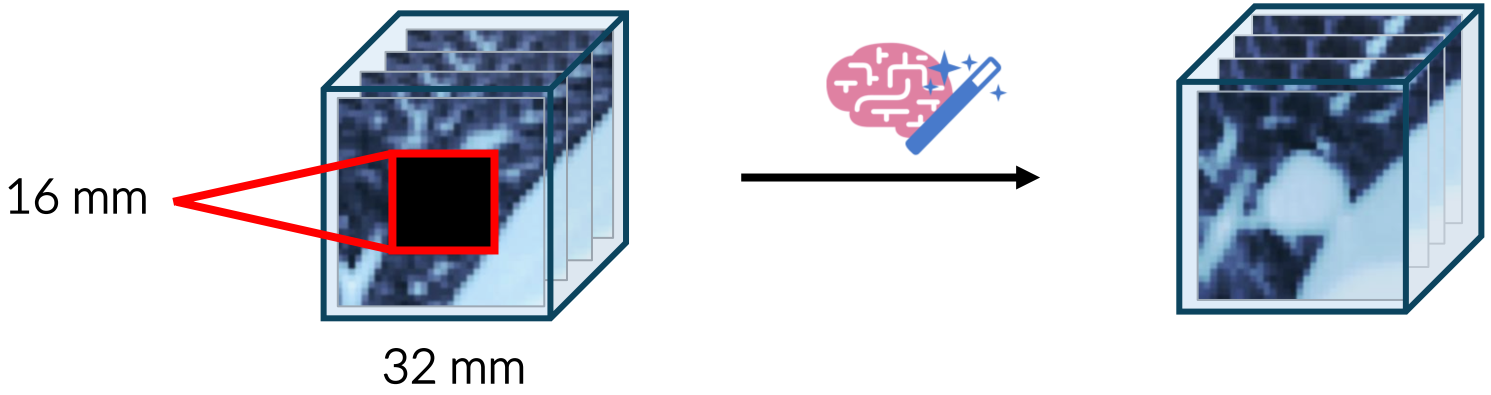


Data generation process

- The central cube of the selected sample is **masked** with zeros and then processed

- The generative network creates the nodule anew

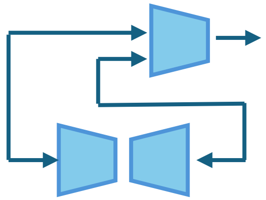
- To preserve the anatomical information the process is **conditioned** with the surrounding pulmonary tissue



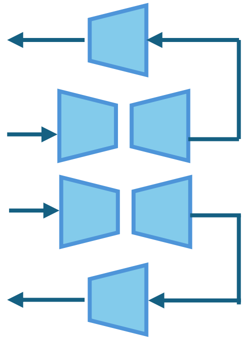
Generative architectures

- We build three versions of the same manipulated CT scan using different generative methods
- We consider Generative Adversarial Networks (GAN) and Diffusion Models (DM)

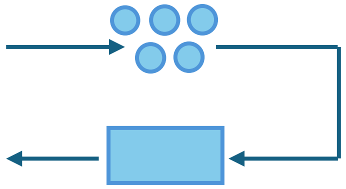
Pix2Pix GAN



CycleGAN

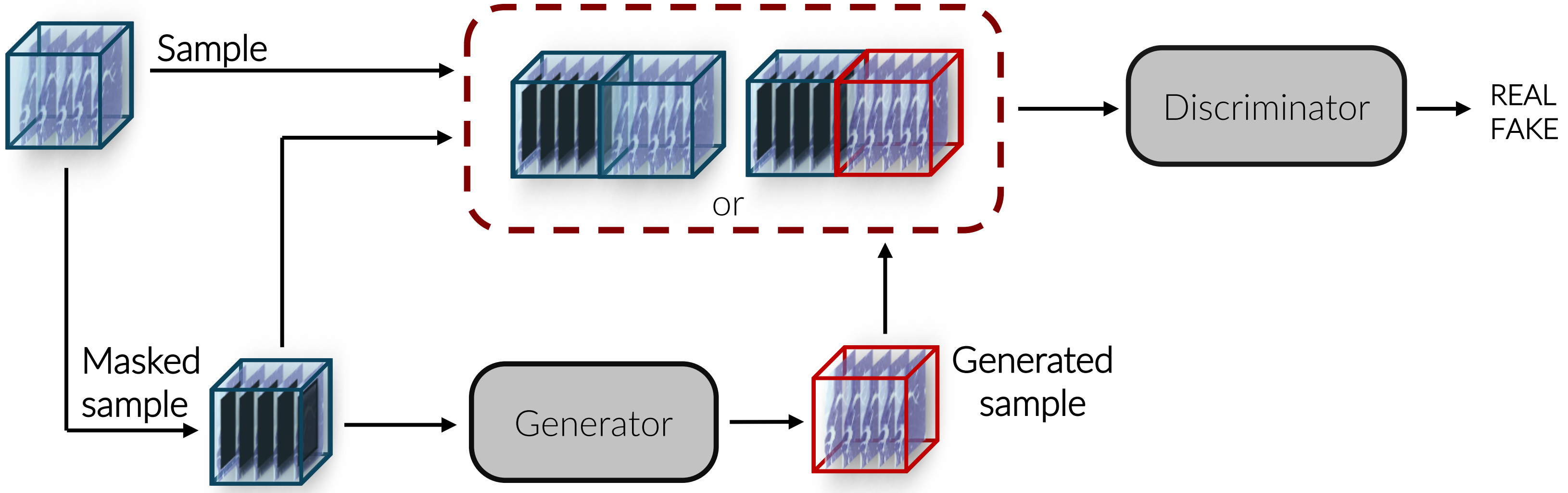


Diffusion Model



Generative architecture: Pix2Pix GAN

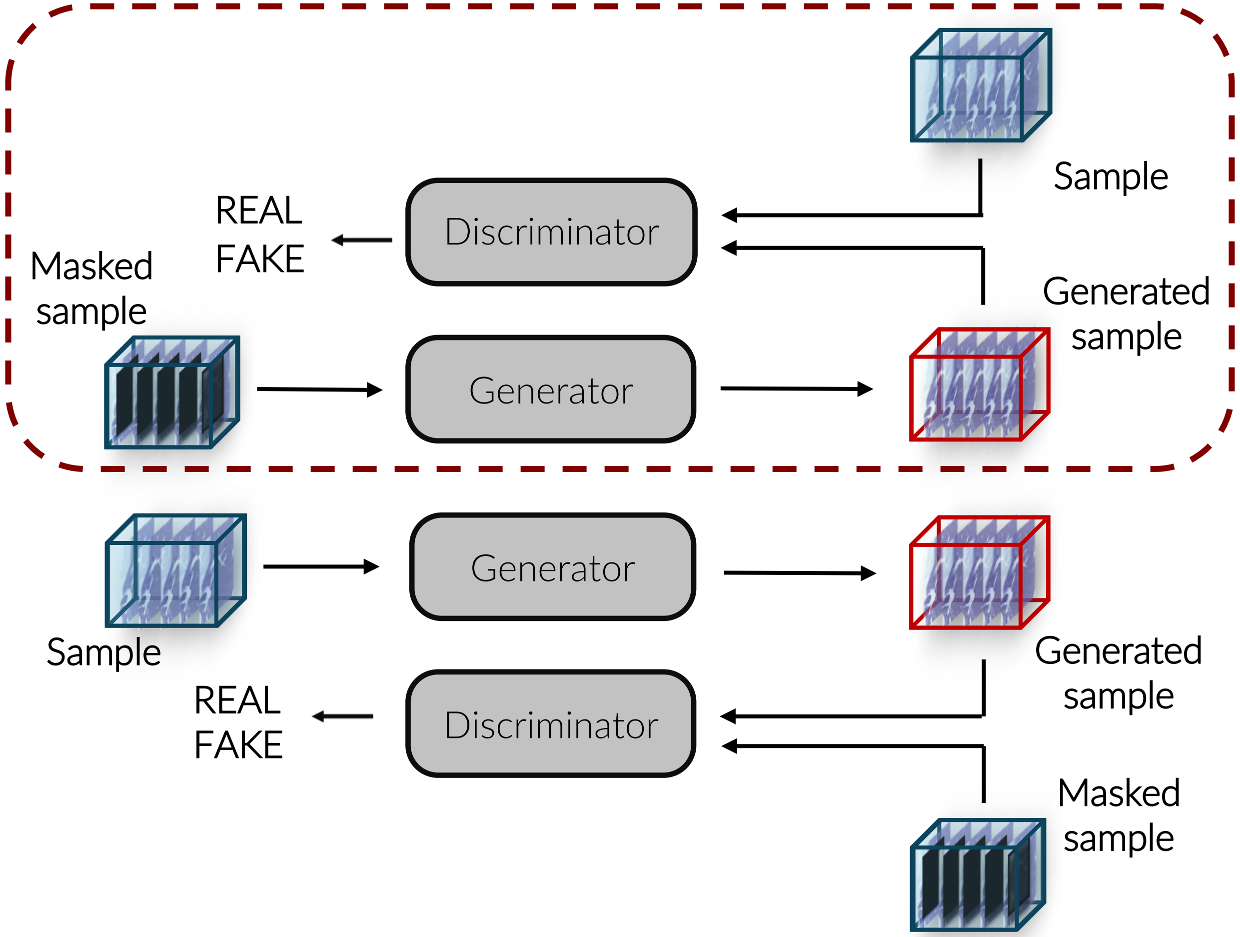
- This is the 3D version of the conditional generative network Pix2Pix GAN [3,4]
- The masked cube guides the process since the generated cube has to be coherent with the original sample



[3] Y. Mirsky et al. "CT-GAN: Malicious tampering of 3d medical imagery using deep learning," in 28th USENIX Security Symposium, 2019.
[4] P. Isola et al. "Image-toimage translation with conditional adversarial networks" CVPR 2017.

Generative architecture: CycleGAN

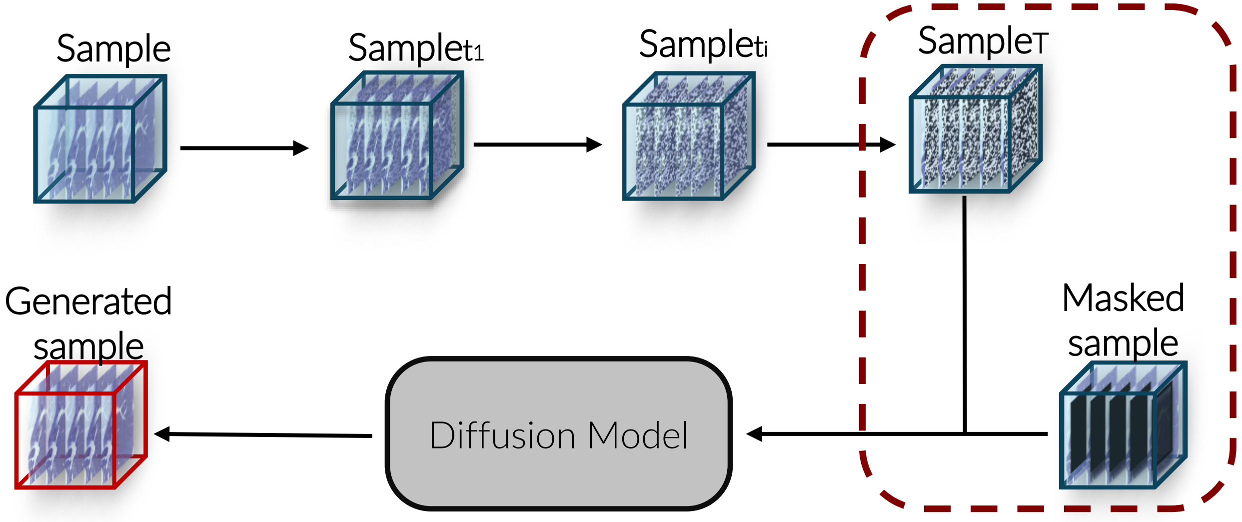
- It is based on the 3D CycleGAN [5], adapted to operate on 3D cubes
- We consider only the translation from masked cubes to synthetic cancerous/non-cancerous tissue



[5] D. Iommi, 3D-CycleGan-Pytorch-Medical-Imaging-Translation, https://github.com/davidiommi/ 3D-CycleGan-Pytorch-MedImaging

Generative architecture: Diffusion Model

- The model is based on the Denoising Diffusion Probabilistic Model [6] adapted for medical images [7]
- To perform the inpainting task the denoiser is provided with an additional input set to the masked cube



[6] J. Ho et al. "Denoising diffusion probabilistic models" NeurIPS 2020
[7] Z. Dorjsembe et al. "Threedimensional medical image synthesis with denoising diffusion probabilistic models," in MIDL 2022

Qualitative analysis

- Evaluation of the generated images through a **computer-aided diagnostic tool** [8]
- The tool localizes the nodules and provides a score of their potential cancerous condition
- The network is applied at the position where the nodule was **injected or removed**

INJECTED NODULE

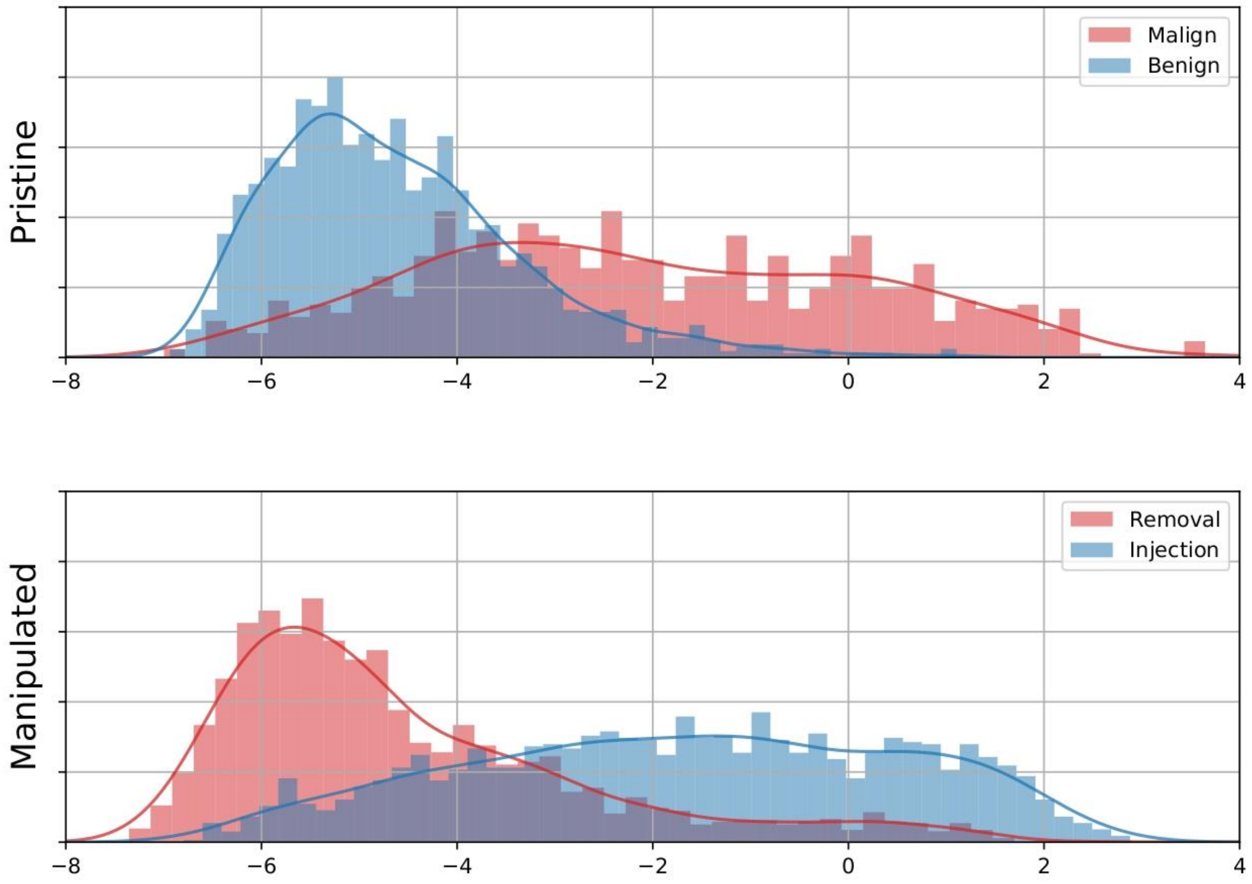
MALIGNACY SCORE: 0.80

[8] F. Liao et al. "Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network" IEEE TNNLS 2019

Qualitative analysis

- The diagnostic tool provides **inverted diagnosis**
- The **removed nodules** have the same histogram as **pristine benign nodules**
- The **injected nodules** are classified as malignant nodules, showing a similar trend to the **pristine malignant ones**

Histrograms of the pristine and manipulated scans



Benchmark: preliminary experiment

- The **forensics detector** [9] trained on general purpose (G.P.) images fails on M3Dsynth images
- The method has no clue on the nature of the medical images

		Test Set					
		General purpose images			M3Dsynth		
	Training Set	ProGAN	StyleGAN2	LDM	Pix2Pix	CycleGAN	DM
G. P images	ProGAN	99.9	98.1	57.1	$\boxed{50.0}$	$\boxed{47.1}$	$\boxed{48.8}$
	StyleGAN2	99.9	100	57.9	$\boxed{50.4}$	$\boxed{49.6}$	$\boxed{52.0}$
	LDM	50.8	50.0	100	$\boxed{44.6}$	$\boxed{44.5}$	$\boxed{46.2}$
M3Dsynth	Pix2Pix	50.5	49.0	48.9	99.5	96.6	95.8
	CycleGAN	49.5	49.0	49.9	97.7	98.5	91.6
	DM	50.9	50.6	50.7	96.1	92.8	97.3

[9] R. Corvi, et al. "On the detection of synthetic images generated by diffusion models," in IEEE ICASSP 2023.

Benchmark: preliminary experiment

- The **forensics detector** [9] trained on general purpose (G.P.) images fails on M3Dsynth images
- The method has no clue on the nature of the medical images

		Test Set					
		General purpose images			M3Dsynth		
	Training Set	ProGAN	StyleGAN2	LDM	Pix2Pix	CycleGAN	DM
G. P images	ProGAN	99.9	98.1	57.1	50.0	47.1	48.8
	StyleGAN2	99.9	100	57.9	50.4	49.6	52.0
	LDM	50.8	50.0	100	44.6	44.5	46.2
M3Dsynth	Pix2Pix	50.5	49.0	48.9	**99.5**	**96.6**	**95.8**
	CycleGAN	49.5	49.0	49.9	**97.7**	**98.5**	**91.6**
	DM	50.9	50.6	50.7	**96.1**	**92.8**	**97.3**

Very different results after fine-tuning

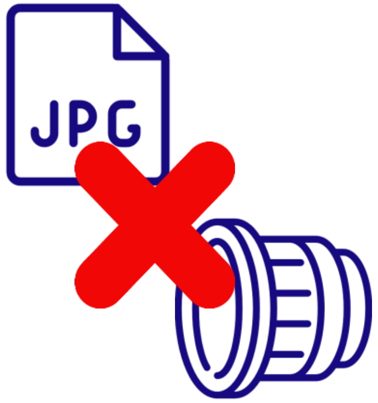
[9] R. Corvi, et al. "On the detection of synthetic images generated by diffusion models," in IEEE ICASSP 2023.

Benchmark: SOTA detectors

There are main differences between medical and general purpose images:

- **Compression techiques** are not customary for CT-scans

- **Medical imaging sensors** have different properties than smartphones or general cameras



Classical approaches which look for compression artifacts or traces of internal camera processing are not suitable for this task

Benchmark: SOTA detectors

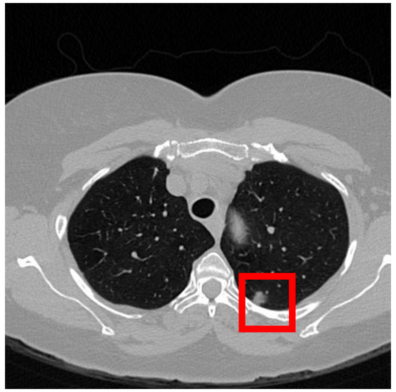
- We choose the following generic forensics methods fine-tuned on **our dataset M3Dsynth**

Method	RGB	Others	Reference
Xception	✓	-	F. Chollet, "Xception: Deep learning with depthwise separable convolutions," CVPR 2017
U-Net	✓	-	O. Ronneberger et al. "U-net: Convolutional networks for biomedical image segmentation" MICCAI 2015.
HP-FCN	-	HP filters	H. Li and J. Huang, "Localization of deep inpainting using high-pass fully convolutional network" ICCV 2019.
ManTraNet	✓	HP filters	Y. Wu et al. "ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features" CVPR 2019.
MVSS-Net	✓	Trainable HP filter	X. Chen et al. "Image Manipulation Detection by Multi-View Multi-Scale Supervision" ICCV 2021.
TruFor	✓	Noiseprint++	F. Guillaro et al. "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," CVPR 2023.

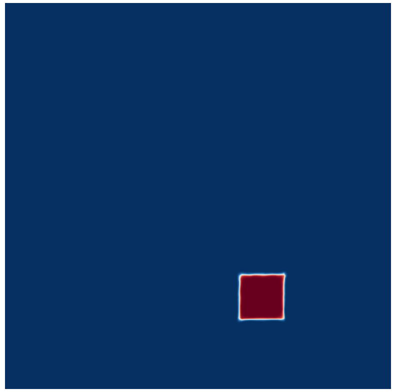
Experimental analysis: metrics

- **Detection:** Pd@1% and balanced accuracy by comparing the maximum detection score obtained over all slices of an image
- **Localization:** F1 measure and IoU metric by comparing the generated 3D localization map and the ground truth

Ground Truth



Localization map



Experimental analysis: results

- **Localization:** the performance is good on average especially for TruFor and ManTraNet

- **Detection:** several methods show good detection performance showing lower results only in few cases (HP-FCN and U-Net)

	Test Set	Pix2Pix			CycleGAN			DM		
	Training Set	Pix2Pix	CycleGAN	DM	Pix2Pix	CycleGAN	DM	Pix2Pix	CycleGAN	DM
F1 / IoU	U-Net [7]	44.5 / 30.7	39.7 / 26.6	35.5 / 23.2	34.4 / 23.3	57.5 / 43.6	22.7 / 15.5	46.9 / 33.3	49.1 / 35.8	57.7 / 43.6
	HP-FCN [8]	85.0 / 75.3	59.1 / 43.4	45.6 / 31.3	63.6 / 49.8	84.5 / 75.3	36.4 / 24.6	77.0 / 64.9	73.6 / 61.9	84.9 / 75.4
	ManTraNet [9]	87.0 / 79.1	66.5 / 50.5	61.4 / 45.5	74.8 / 63.3	85.5 / 77.2	60.5 / 47.4	83.2 / 73.0	81.8 / 70.7	87.2 / 78.5
	MVSS-Net [10]	81.4 / 70.4	63.2 / 49.8	56.8 / 42.5	74.7 / 64.2	86.2 / 78.0	55.1 / 44.1	79.5 / 68.5	72.8 / 62.2	84.9 / 75.4
	TruFor [11]	89.9 / 82.9	68.1 / 55.5	68.0 / 54.7	79.0 / 70.1	88.2 / 81.2	65.0 / 54.1	84.4 / 75.2	76.9 / 66.7	89.3 / 82.0
Acc / Pd@1%	Xception [6]	83.7 / 99.8	86.9 / 95.2	71.9 / 80.3	81.3 / 86.1	87.4 / 99.2	64.1 / 37.8	83.5 / 97.7	86.8 / 94.1	71.9 / 96.9
	U-Net [7]	52.9 / 93.1	60.3 / 74.5	53.7 / 56.5	52.1 / 64.4	60.6 / 95.4	53.0 / 29.2	52.9 / 91.1	60.3 / 79.5	53.7 / 96.8
	HP-FCN [8]	59.8 / 45.6	71.4 / 50.8	60.2 / 31.7	59.8 / 43.1	71.4 / 52.0	60.3 / 28.9	59.8 / 45.4	71.4 / 51.4	60.4 / 33.6
	ManTraNet [9]	52.7 / 100.	56.6 / 99.9	52.8 / 91.2	52.7 / 93.4	56.6 / 99.7	52.8 / 87.3	52.7 / 99.9	56.6 / 100.	52.8 / 100.
	MVSS-Net [10]	73.0 / 95.8	92.5 / 97.2	75.4 / 86.2	72.1 / 70.8	92.7 / 99.3	73.7 / 67.4	73.0 / 91.2	92.6 / 97.9	76.0 / 99.3
	TruFor [11]	95.0 / 100.	95.8 / 97.8	94.3 / 97.0	93.3 / 95.9	96.0 / 99.4	91.2 / 89.1	95.0 / 99.9	96.0 / 98.1	94.9 / 99.6

Experimental analysis: 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			CycleGAN			DM		
	Training Set	Pix2Pix	CycleGAN	DM	Pix2Pix	CycleGAN	DM	Pix2Pix	CycleGAN	DM
F1 / IoU	U-Net [7]	44.5 / 30.7	39.7 / 26.6	35.5 / 23.2	34.4 / 23.3	57.5 / 43.6	22.7 / 15.5	46.9 / 33.3	49.1 / 35.8	57.7 / 43.6
	HP-FCN [8]	85.0 / 75.3	59.1 / 43.4	45.6 / 31.3	63.6 / 49.8	84.5 / 75.3	36.4 / 24.6	77.0 / 64.9	73.6 / 61.9	84.9 / 75.4
	ManTraNet [9]	87.0 / 79.1	66.5 / 50.5	61.4 / 45.5	74.8 / 63.3	85.5 / 77.2	60.5 / 47.4	83.2 / 73.0	81.8 / 70.7	87.2 / 78.5
	MVSS-Net [10]	81.4 / 70.4	63.2 / 49.8	56.8 / 42.5	74.7 / 64.2	86.2 / 78.0	55.1 / 44.1	79.5 / 68.5	72.8 / 62.2	84.9 / 75.4
	TruFor [11]	89.9 / 82.9	68.1 / 55.5	68.0 / 54.7	79.0 / 70.1	88.2 / 81.2	65.0 / 54.1	84.4 / 75.2	76.9 / 66.7	89.3 / 82.0
Acc / Pd@1%	Xception [6]	83.7 / 99.8	86.9 / 95.2	71.9 / 80.3	81.3 / 86.1	87.4 / 99.2	64.1 / 37.8	83.5 / 97.7	86.8 / 94.1	71.9 / 96.9
	U-Net [7]	52.9 / 93.1	60.3 / 74.5	53.7 / 56.5	52.1 / 64.4	60.6 / 95.4	53.0 / 29.2	52.9 / 91.1	60.3 / 79.5	53.7 / 96.8
	HP-FCN [8]	59.8 / 45.6	71.4 / 50.8	60.2 / 31.7	59.8 / 43.1	71.4 / 52.0	60.3 / 28.9	59.8 / 45.4	71.4 / 51.4	60.4 / 33.6
	ManTraNet [9]	52.7 / 100.	56.6 / 99.9	52.8 / 91.2	52.7 / 93.4	56.6 / 99.7	52.8 / 87.3	52.7 / 99.9	56.6 / 100.	52.8 / 100.
	MVSS-Net [10]	73.0 / 95.8	92.5 / 97.2	75.4 / 86.2	72.1 / 70.8	92.7 / 99.3	73.7 / 67.4	73.0 / 91.2	92.6 / 97.9	76.0 / 99.3
	TruFor [11]	95.0 / 100.	95.8 / 97.8	94.3 / 97.0	93.3 / 95.9	96.0 / 99.4	91.2 / 89.1	95.0 / 99.9	96.0 / 98.1	94.9 / 99.6

Conclusions

- We introduced M3Dsynth a new large dataset of tampered 3D medical images with local AI-based manipulations

- The dataset has been used to train and test several state of-the-art methods which proved good both at detecting and localizing local manipulations

- Despite the good results we believe that with new and more sophisticated AI-generative techniques, it would be important to develop forensic approaches specifically tailored to medical data

Conclusions

- We introduced M3Dsynth a new large dataset of tampered 3D medical images with local AI-based manipulations
- The dataset has been used to train and test several state of-the-art methods which proved good both at detecting and localizing local manipulations
- Despite the good results we believe that with new and more sophisticated AI-generative techniques, it would be important to develop forensic approaches specifically tailored to medical data

Any questions?