MAM-E: Mammographic synthetic image generation with diffusion models

Ricardo Montoya del Ángel

Supervisor: Robert Martí Marly

Computer vision and robotics institute (ViCOROB)
Universitat de Girona

July 4, 2023

Motivation

State-of-the-art of generative models

Text generation

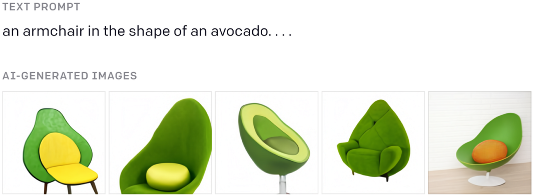
ChatGPT

Code generation

GitHub Copilot

State-of-the-art of generative models

Image generation



The problem with medical images...

Data scarcity main reasons:

- Intrinsically more expensive due to their acquisition, processing and labeling procedure.
- More privacy and data protections
- Some rare cases are difficult to find

Consequence

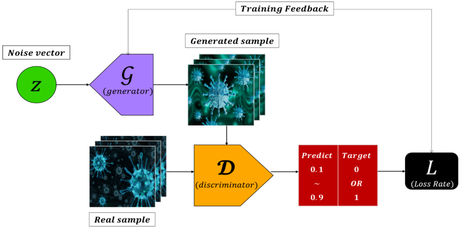
This limits the performance of deep learning models and holds back the development of CAD systems.

Solution: generate new data

Generative models have been used to complement traditional data augmentation techniques.

Generative adversarial networks (GANs)

For many years the state-of-the-art (SOTA) for synthetic image generation tasks.



Yann LeCun on GANs (Barcelona, 2016):

"..the most interesting idea in the last 10 years."

Diffusion models

In 2021 OpenAI's belligerent article appears [1].

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Diffusion models strong points over GANs:

- Image sample quality
- Generation diversity (mode collapse problem)
- Training stability (convergence problem)

Diffusion in medical imaging

- Diffusion models frameworks started to be used in the medical domain.
- **Still to this day** there is no implementation of diffusion model techniques for mammography synthesis.

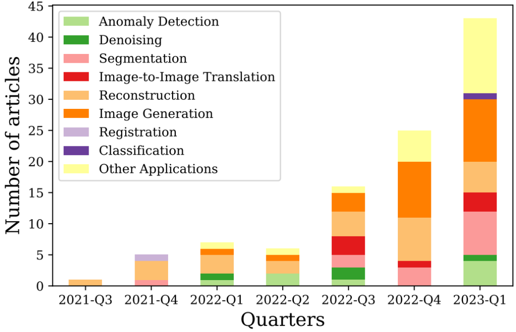


Figure: From Kazerouni et al. (2023) [2]

Purpose of this work

Thesis objectives

- **Explore** the use of diffusion models for the generation of high-resolution mammographic images and **develop** a synthesis pipeline using SOTA conditional diffusion models to control the image generation process.

Task 1: Conditional generation

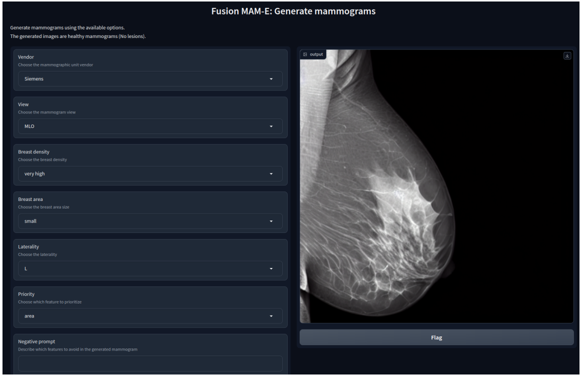
Generate **healthy mammograms** based on a text prompt controlling:

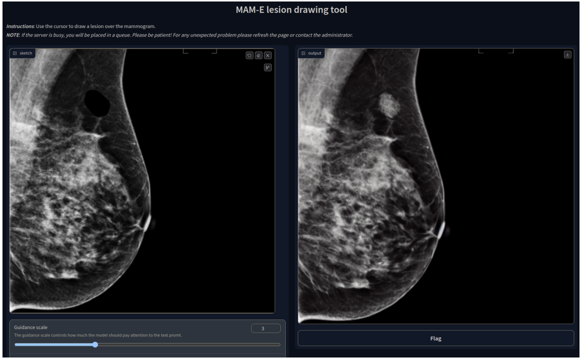
- View
- Breast density
- Breast area
- Vendor

Task 2: Lesion inpainting

Generate synthetic lesions in desired regions based on a mask or region of interest (ROI).

We present *MAM-E*





Methods

Datasets

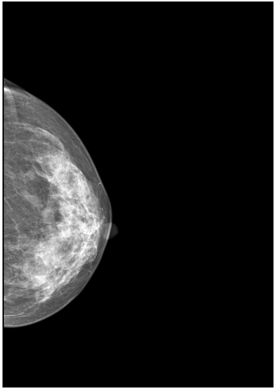
OMI-H (UK)

- Subset of the OPTIMAM Mammography Image Database
- Around 40k Full Field digital mammograms (FFDM)
- Hologic vendor
- Healthy patients and with lesions
- CC, MLO, ML and LM views.
- Metadata at patient and image level

Datasets

VinDr-mammo (Vietnam)

- Around 20k Full Field digital mammograms (FFDM)
- Siemens vendor
- Mammograms with and without lesions
- Only CC and MLO views
- Metadata with clinical information per image



Datasets stats

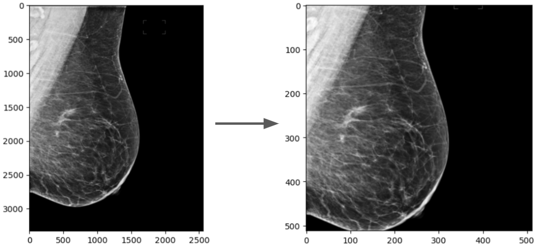
Table: Distribution of cases for both datasets.

	OMI-H	VinDr	Combined
Healthy	33,643	13,942	47,585
With lesion	6,908	1,533	8,441
Total	40,551	15,475	56,026

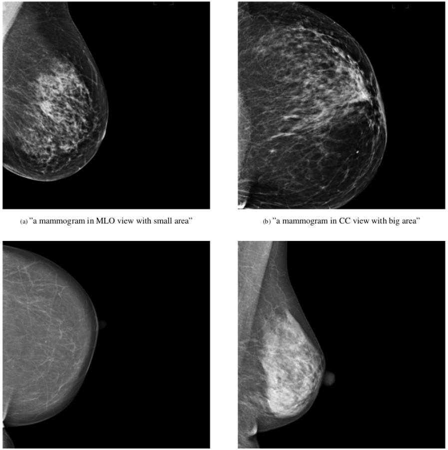
Preprocessing

For both datasets

- From DICOM to PNG
- Stored in RGB format (Repeating gray channel)
- Data type from unsigned 16 bits to 8 bits.
- Cropping and resize for a 512x512 resolution
- Images with R laterality are flipped.



Data preparation for conditional model (task 1)



(a) "a mammogram in MLO view with small area"

(b) "a mammogram in CC view with big area"

(c) "a mammogram in CC view with very low density"

(d) "a mammogram in MLO view with very high density"

Image-prompt pair

Every **healthy** mammogram is paired with a text description.

- View
- Breast area size (OMI-H)
- Breast density (VinDr)

Data preparation for inpainting model (task 2)

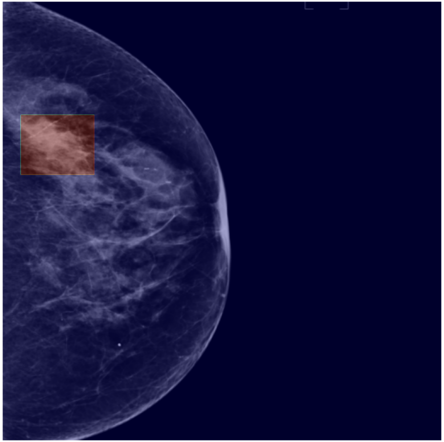


Image-mask pair

Every mammogram with lesion is paired with a rectangular mask containing the lesion. The mask is a binary image with 255 value inside the lesion and 0 elsewhere.

What are diffusion models?

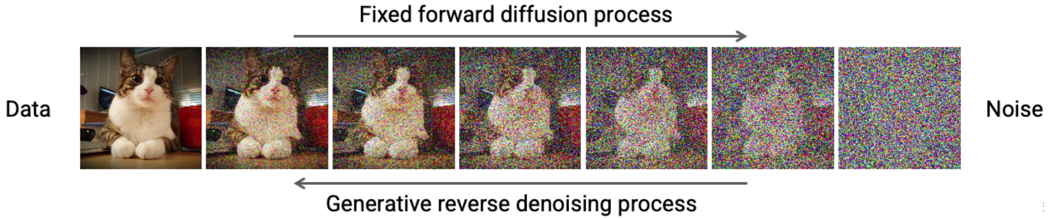
- Original idea was presented in 2015 by Sohl-Dickstein et al. [3], inspired in non-equilibrium thermodynamics.

Diffusion models main idea

Data distributions can be learned by:

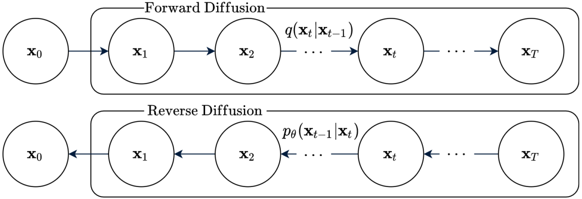
$\rightarrow$ **Systematically and iteratively** destroy input information by adding certain noise.

$\leftarrow$ Tasking a neural network to learn to remove it in a denoising process.



Denoising diffusion probabilistic models (DDPM)

In the first practical implementation by Ho et al. (2020) [4]:



- Adding Gaussian noise to the image.
- Total number of diffusion timesteps $T \approx 1k$
- A encoder-decoder neural network (like a UNet) is used for the denoising process.

Forward diffusion

Adding noise to an image x_0

$$x_t = \sqrt{\tilde{\alpha}_t} x_0 + \sqrt{1 - \tilde{\alpha}_t}\epsilon. \tag{1}$$

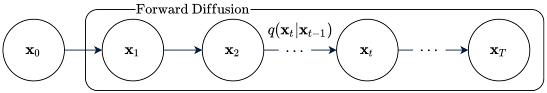
where $\epsilon \sim N(0, 1)$.

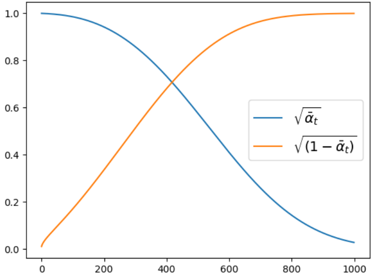
Forward diffusion

Adding noise to an image x_0

$$x_t = \sqrt{\tilde{\alpha}_t} x_0 + \sqrt{1 - \tilde{\alpha}_t}\epsilon. \tag{1}$$

where $\epsilon \sim N(0, 1)$.





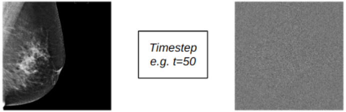
Reverse diffusion

Simplified optimization term

$$L = \|\epsilon - \epsilon_\theta\|^2, \quad (2)$$

where $\epsilon \sim N(0, 1)$.

Training a diffusion model



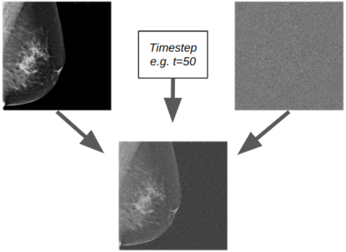
Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Training a diffusion model

Algorithm 1 Training

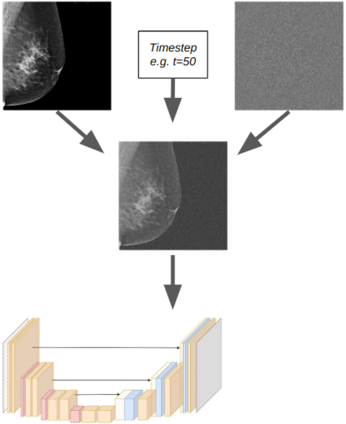
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged



Training a diffusion model

Algorithm 1 Training

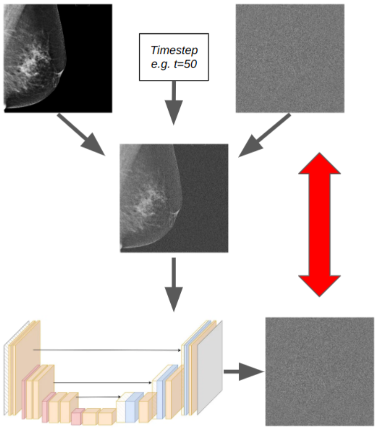
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged



Training a diffusion model

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged



Vanilla diffusion

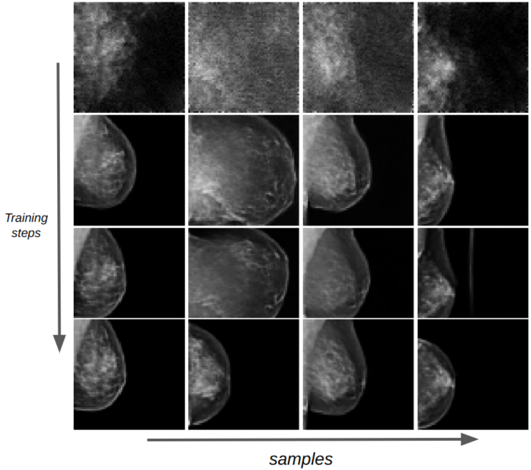


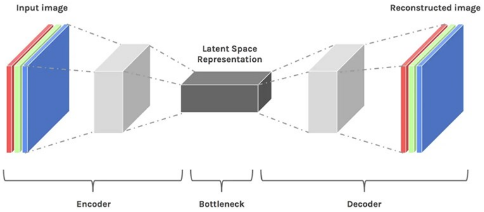
Figure: Training evolution of vanilla diffusion model at epoch 1, 16, 30 and 50.

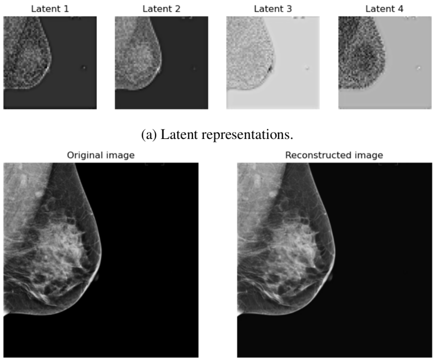
Latent diffusion

Use encoder to **compress** the image into a smaller representation and apply diffusion on the **latent space**.

Variational Autoencoder (VAE)

Converts an original 512x512 RGB image into a 64x64x4 latent representation.





(a) Latent representations.

(b) The original image and its reconstruction.

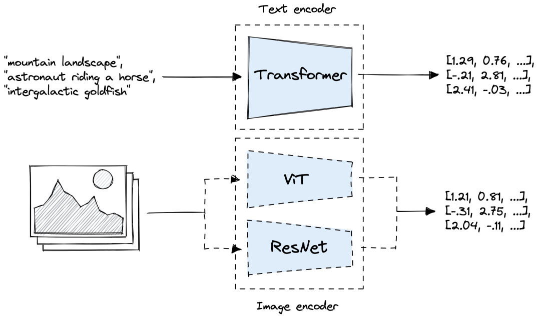
Stable diffusion

Improvement of Latent diffusion

- Text conditioning is added to the model for additional control on the generation process.
- A **numeric representation of the prompt** is created using a pretrained transformer called CLIP [5].

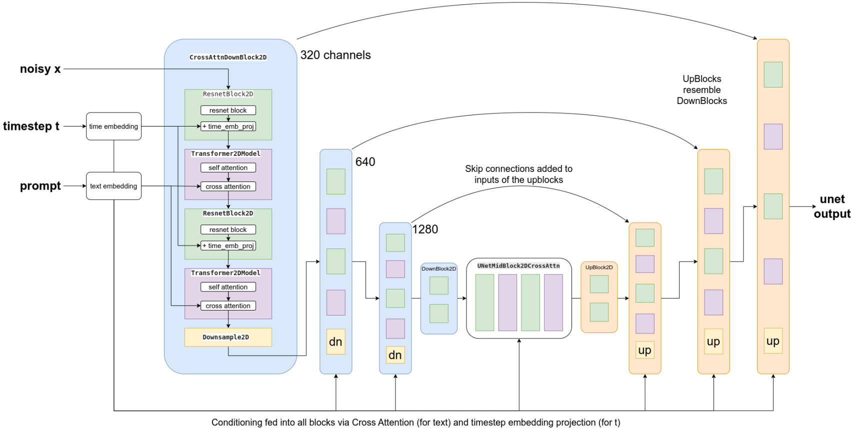
Contrastive Language-Image Pre-training (CLIP)

It can map both text and images into the same representational space, **allowing comparison** and similarity quantification.



Stable diffusion

The CLIP embedings are used in the attention layers of the UNet through a **cross-attention mechanism**.



Fine-tuning SD: DreamBooth

- In 2022 Stability AI and LAION made the **pretrained weights** of the Stable diffusion pipeline [6] publicly available.
- In the same year Ruiz et al. [7] proposed the DreamBooth fine-tuning technique.

DreamBooth characteristics

- Use only a few images of the new subject with its respective text prompt.
- Train using a small learning rate.
- Decide which models to train: UNet, VAE and CLIP.

DreamBooth allows:

- Binding of the new subject to a **new unique identifier** in the text embedding space.
- A **new learned representation** in the pretrained UNet data distribution.

Healthy mammogram synthesis

Our proposal

- Adapt the DreamBooth technique to our synthetic mammography generation task.
- Train a model for each dataset and a combination of both.
- Freeze VAE. Train UNet and CLIP together.

Training Hyperparameters

- Batch size: We ranged from 8, 16, 32, 64, 128 and 256.
- Training steps: Experiments ranged from 1k up to 16k.
- Learning rate: We explore three main values $1e^{-6}$, $1e^{-5}$, $1e^{-4}$.
- Gradient clipping normalization: 1.0
- Optimizer: Adam weight decay

Healthy mammogram synthesis

Our proposal

- Adapt the DreamBooth technique to our synthetic mammography generation task.
- Train a model for each dataset and a combination of both.
- Freeze VAE. Train UNet and CLIP together.

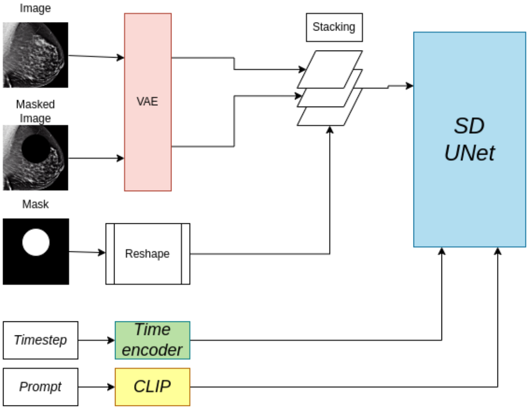
Tracking training performance

- 4 sample images are generated every 1k or 2k training steps from the same seed. (Same Gaussian noise)
- Training loss is also monitored.
- Training logs, config file and all relevant files uploaded in the Weights and Biases cloud (Publicly available).

Lesion inpainting

Our proposal

- Modify key aspects of the SD pipeline to perform the inpainting task.
- Two new elements are added: the mask and a masked version of the original image.
- The latent representation, the mask, and the masked latent representation are stacked into one tensor
- This allows the network to pay attention only to the pixels inside the mask.



Resources management

Loading 3 models to a GPU and enabling the gradient tracking for two of them for training requires **large GPU memory**.

GPU memory usage reduction techniques

- Mixed precision using the *fp16* arithmetic.
- 8-bit AdamW optimizer by *BitsAndBytes*.
- *Xformers* efficient memory usage for transformers.
- Gradient accumulation.
- Gradient checkpointing.
- Set the optimizer gradients to None instead of zero after the weights update.

This reduces the GPU memory to <u>only 20 GB</u>.

Results

Qualitative assessment

Hologic mammograms model

Prompt: "a mammogram in MLO view with small area".

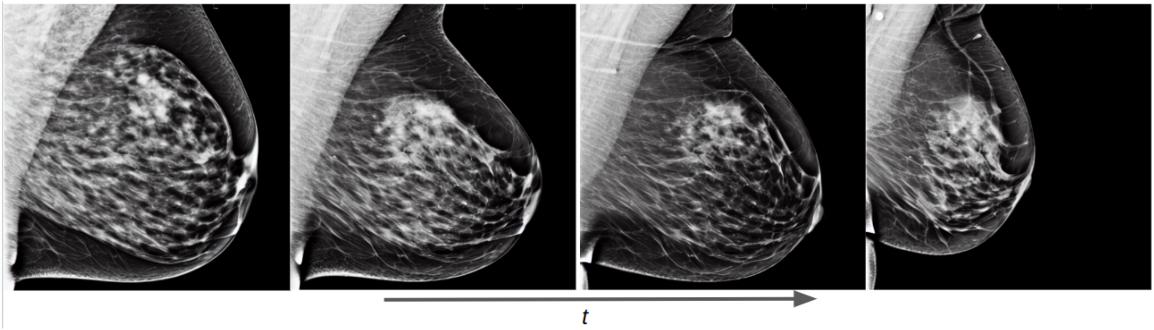


Figure: Denoising process at epoch 1, 3, 6 and 10.

Qualitative assessment

Siemens mammograms model

Prompt: "a mammogram in CC view with high density".

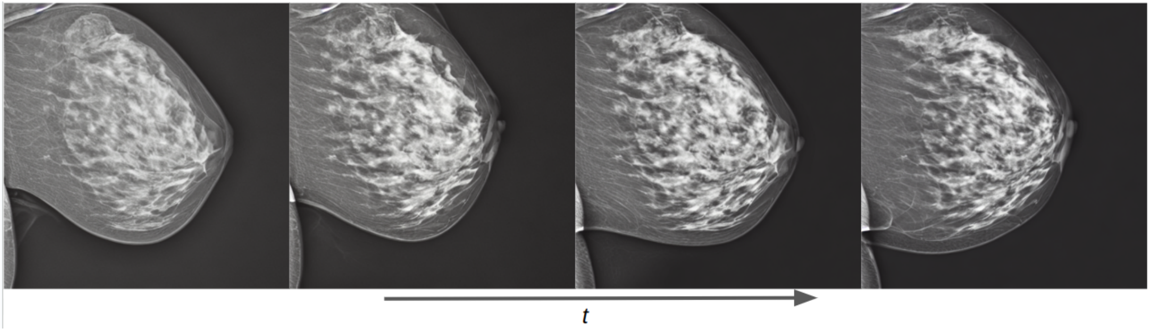


Figure: Denoising process at epoch 1, 3, 6 and 10.

Qualitative assessment

Combined mammograms (fusion) model

Prompt: "a siemens mammogram in MLO view with high density and small area".

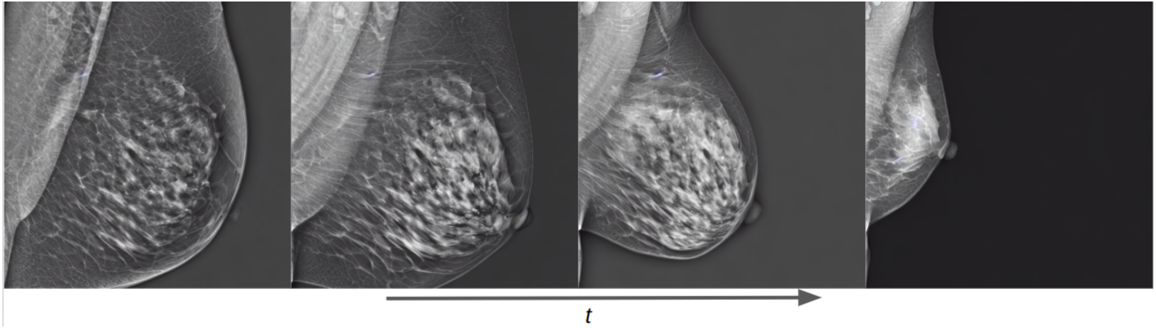
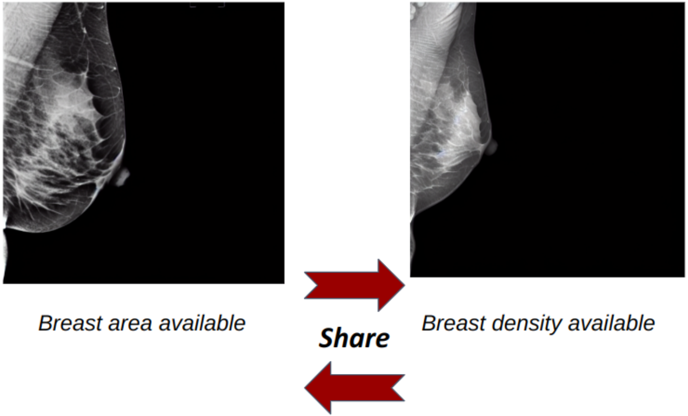


Figure: Denoising process at epoch 1, 3, 7 and 40.

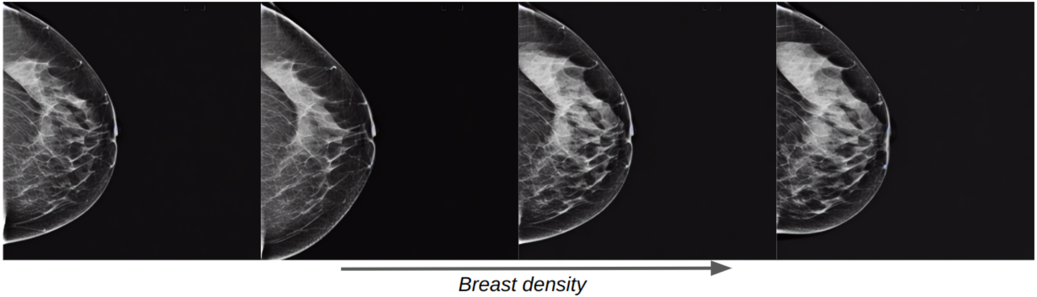
Concept extrapolation

The fusion model allows to **extrapolate** the characteristics of one dataset to the other.



Concept extrapolation

The fusion model allows to **extrapolate** the characteristics of one dataset to the other.



Radiological assessment

Quality test

We asked a radiologist with 30 years of experience to assess 53 real and synthetic images. The criteria were:

- 0: Definitely real
- 1: Probably real
- 2: Not sure
- 3: Probably synthetic
- 4: Definitely synthetic.

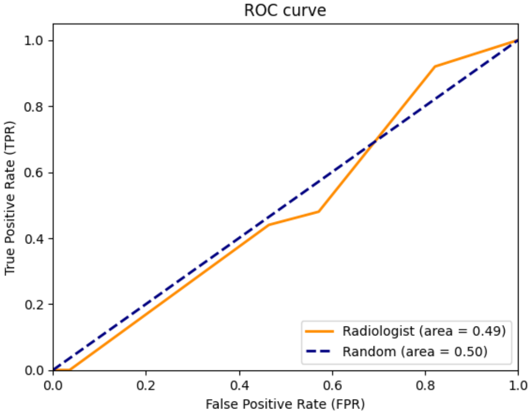


Figure: ROC curve of radiological assessment.

Inference: image generation

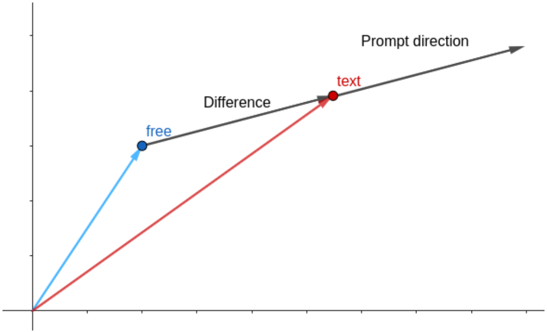
Two main inference parameters

Denoising timsteps T

- Depends on the sampling method.
- The DPM-solver [8] allows fast diffusion sampling with only 20 steps.

Guidance scale

- Classifier-free guidance parameter to control the "importance" of the text prompt.
- Values usually range around 4-7.



$$\epsilon_\theta = \epsilon_{free} + guidance * (\epsilon_{text} - \epsilon_{free}). \quad (3)$$

Quantitative assessment

Guidance effect

The value of the guidance scale directly affects both prompt **fidelity** and generation **diversity**.

- A very small guidance creates a low quality and low fidelity image.

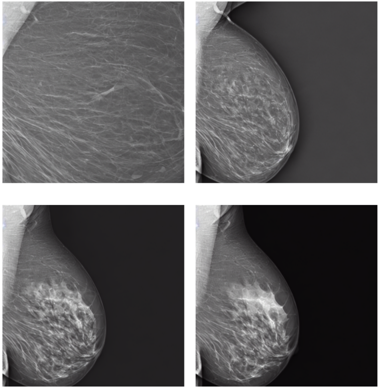


Figure: Guidance scale from 1 to 4.

Quantitative assessment

Guidance effect

The value of the guidance scale directly affects both prompt **fidelity** and generation **diversity**.

- A very small guidance creates a low quality and low fidelity image.
- A very large guidance reduces the generation diversity.

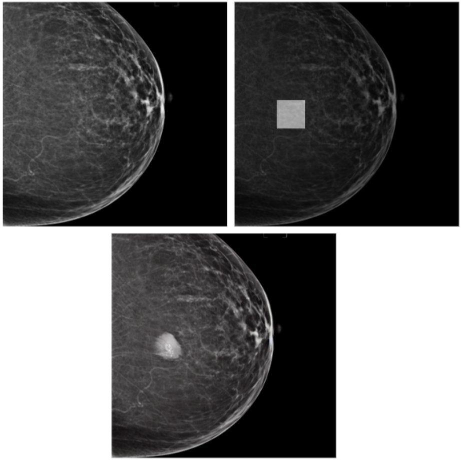
Figure: MS-SSMI value between 50 generated images.

	Hologic		Siemens	
Guidance	Mean↓	STD	Mean↓	STD
4	0.29	0.16	0.38	0.19
5	0.34	0.16	0.36	0.17
6	0.38	0.12	0.41	0.17
7	0.38	0.1	0.34	0.17
8	0.43	0.11	0.42	0.2
9	0.42	0.13	0.43	0.16
10	0.49	0.12	0.41	0.13
11	0.5	0.12	0.47	0.17
12	0.52	0.11	0.46	0.16
13	0.48	0.1	0.42	0.16
14	0.5	0.11	0.4	0.18

Lesion inpainting

Main results

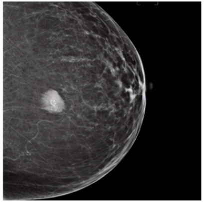
- A synthetic lesion can be inpainted over a mammogram, given a lesion bounding box (mask).



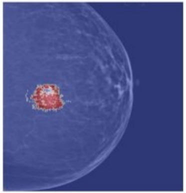
Lesion inpainting

Main results

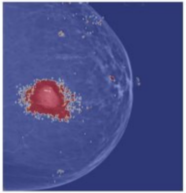
- A synthetic lesion can be inpainted over a mammogram, given a lesion bounding box (mask).
- 3 Explainability AI methods show sensibility of synthetic lesions to a lesion classification CAD system (Sam, MAIA 2023).



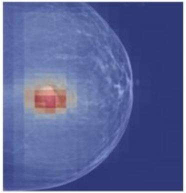
(a) Inpainted synthetic lesion.



(b) Gradcam heatmap.



(c) Saliency heatmap.

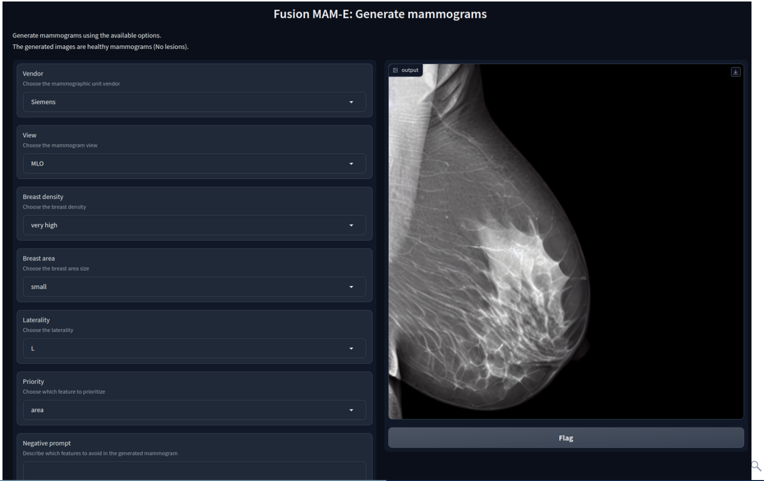


(d) Occlusion heatmap.

Graphical user interface

Fusion **MAM-E**

- 3 different GUIs, one for each dataset combination.
- Extra feature: negative prompt
- Designed using *GradIO*, an open-source Python package for rapid generation of visual interface of ML models.



Conclusions

Conclusions

We accomplish:

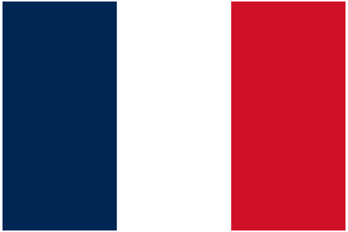
- Implementation of the SD pipeline for synthetic mammography generation.
- Fine-tuning of a SD model pretrained with **natural images** using the DreamBooth technique.
- Synthesis of mammograms with specific characteristics controlled by a text prompt and **concept extrapolation**.
- Modification of SD for **inpainting synthetic lesions** over healthy mammograms.
- First work to use SD fine-tuning for lesion inpainting in medical images and first use of DM for mammograms.
- Initial evidence that synthetic images coming from our implementation could potentially be used for CAD systems.

Limitations and future work

- Reduced data type (16 bits $\rightarrow$ 8 bits)
- Limited resolution (512x512) and FOV.
  - Newer pretrained weights for higher resolution
  - Super-resolution models
  - Keep original rectangular ratio
- Quantitative assessment at training and inference time.
- Absence of complete CAD pipelines training with synthetic images to analyze performance changes.
- Extensive hyperparameter exploration.

Grazie!

Gràcies!



Merci!

References

References I

[1] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," Jun. 2021, arXiv:2105.05233 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2105.05233

[2] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, Aug. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841523001068

[3] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," Nov. 2015, arXiv:1503.03585 [cond-mat, q-bio, stat]. [Online]. Available: http://arxiv.org/abs/1503.03585

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Dec. 2020, arXiv:2006.11239 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.11239

References II

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 2022, arXiv:2112.10752 [cs]. [Online]. Available: http://arxiv.org/abs/2112.10752

[7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," Mar. 2023, arXiv:2208.12242 [cs]. [Online]. Available: http://arxiv.org/abs/2208.12242

References III

[8] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps," Oct. 2022, arXiv:2206.00927 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2206.00927