UNIVERSIDADE DE SÃO PAULO

Escola de Engenharia de São Carlos

Model-based deep learning to restore low-dose digital breast tomosynthesis images

Rodrigo de Barros Vimieiro

EESC · USP

UNIVERSITY OF SÃO PAULO
SÃO CARLOS SCHOOL OF ENGINEERING

Rodrigo de Barros Vimieiro

Model-based deep learning to restore low-dose digital breast tomosynthesis images

São Carlos

2023

Rodrigo de Barros Vimieiro

Model-based deep learning to restore low-dose digital breast tomosynthesis images

Thesis presented to the São Carlos School of Engineering of the University of São Paulo, in partial fulfillment of the requirements for the degree of Doctor of Science - Graduate Program in Electrical Engineering.

Subject area: Signal Processing and Instrumentation

Advisor: Prof. Dr. Marcelo Andrade da Costa Vieira
Co-advisor: Dr. Lucas Rodrigues Borges

VERSÃO CORRIGIDA

São Carlos
2023

Vimieiro, Rodrigo de Barros

S856m Title: Model-based deep learning to restore low-dose digital breast tomosynthesis images / Rodrigo de Barros Vimieiro ; advisor Marcelo Andrade da Costa Vieira ; co-advisor Lucas Rodrigues Borges. – São Carlos, 2023.

Thesis - Graduate Program in Electrical Engineering and Research area in Signal Processing and Instrumentation – São Carlos School of Engineering, University of São Paulo, 2023.

1. Digital breast tomosynthesis 2. Convolutional neural networks 3. Deep Learning 4. Artificial neural networks 5. Model-based deep learning 5. Image restoration I. Model-based deep learning to restore low-dose digital breast tomosynthesis images.

FOLHA DE JULGAMENTO

Candidato: Bacharel **RODRIGO DE BARROS VIMIEIRO.**

Título da tese: "Aprendizagem profunda baseada em modelo para restauração de imagens de tomossíntese digital mamária de baixa dose".

Data da defesa: 06/11/2023.

Comissão Julgadora	**Resultado**
Prof. Associado **Marcelo Andrade da Costa Vieira** **(Orientador)** (Escola de Engenharia de São Carlos/EESC-USP)	APROVADO
Prof. Dr. **César Henrique Comin** (Universidade Federal de São Carlos - UFSCar)	APROVADO
Prof. Dr. **Andrew Douglas Arnold Maidment** (University of Pennsylvania)	APPROVED
Profa. Titular **Agma Juci Machado Traina** (Instituto de Ciências Matemáticas e de Computação - ICMC)	APROVADO
Prof. Dr. **Ge Wang** (Rensselaer Polytechnic Institute, New York)	APROVADO

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **Marcelo Andrade da Costa Vieira**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Carlos De Marqui Junior**

Este trabalho é dedicado a Deus
e aos meus pais, Erika e José Ronaldo.

ACKNOWLEDGMENTS

I want to thank the Rensselaer Polytechnic Institute (RPI), especially Dr. Wang and his group, for having me in the laboratory and sharing all the knowledge.

I want to thank the University of Pennsylvania, especially Dr. Maidment and his group, for their close collaboration.

I also want to thank Real Time Tomography and the whole group, especially Susan, for all the great opportunities.

AGRADECIMENTOS

Este é o capítulo mais importante deste trabalho. O mesmo só foi possível mediante colaboração de inúmeras pessoas e com toda certeza não conseguirei expressar com palavras minha gratidão por todos.

Primeiramente a Deus e a Jesus pela grande oportunidade de desenvolvimento intelectual e também pessoal ao longo de todo o processo.

A toda a minha família sem exceção alguma. Em especial a minha mãe Erika, meu pai José Ronaldo e meu irmão Junior. Ressaltando ainda a dedicação à minha mãe, que luta contra o câncer de mama.

A todos os meus amigos pelo encorajamento, força e apoio.

A Débora, por todo tempo que estivemos juntos e por todo apoio durante esse tempo.

Aos companheiros da república, onde passei grande parte do curso, pelo apoio e companheirismo diário.

Aos grandes amigos da Casa do Caminho de São Carlos, pelo apoio espiritual e pessoal.

A todos os membros e ex-membros do laboratório que ao compartilhar as informações, enriquecem o trabalho de todos. O meu agradecimento a cada um pela enorme contribuição tanto pessoal quanto profissional.

Especialmente ao Prof. Marcelo, pelo acolhimento paternal desde o início dos meus estudos no laboratório, por transmitir-me os conhecimentos e por todo apoio pessoal.

Da mesma forma ao Lucas Borges, por todo o apoio intelectual na pesquisa, paciência no dia a dia e também pelo apoio pessoal.

Indubitavelmente, este trabalho só foi possível de ser feito e concluído mediante a excepcional orientação que tive o privilégio de ter de ambos mencionados acima.

Aos companheiros do Hospital de Amor de Barretos e aos companheiros do Instituto de Radiologia (InRad) do Hospital das Clínicas FMUSP por compartilharem inúmeras imagens clínicas conosco.

A Universidade de São Paulo, a todos os seus colaboradores sem exceção alguma.

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Por fim, que este singelo trabalho possa ser uma pequena contribuição para a grande luta contra o câncer de mama.

“É que existem Espíritos esclarecidos e Espíritos evangelizados, e eu, agora, peço a Deus que abençoe a minha esperança de pertencer ao número destes últimos.”

Francisco Cândido Xavier ditado pelo espírito Humberto de Campos,
Prefácio, Na Escola do Evangelho, Boa Nova.

ABSTRACT

VIMIEIRO, Rodrigo de Barros **Model-based deep learning to restore low-dose digital breast tomosynthesis images**. 2023. 144p. Thesis - São Carlos School of Engineering, University of São Paulo, São Carlos, 2023.

Digital breast tomosynthesis (DBT) and full-field digital mammography (FFDM) are the most commonly used exams for breast cancer screening. In these systems, achieving high image quality is crucial for radiologists to detect the earliest signs of breast cancer and improve the patient's prognosis. The radiation dose is also a concern, given that these systems employ ionizing radiation. While current systems operate within safe radiation margins, there is a growing desire to minimize radiation dose without compromising image quality. To address this challenge, image restoration techniques have emerged as valuable tools to enhance image quality from low-dose (LD) acquisitions. Traditional restoration methods rely on mathematical models that represent the underlying physics of the acquisition system. Convolutional neural networks (CNN), employing modern deep learning (DL) techniques, are capable of learning the image restoration task from data and have exhibited substantial progress in recent years. This work proposes a hybrid model-based deep learning (MBDL) framework for the restoration of DBT images acquired with reduced radiation doses, benefiting from the advantages of both fields. Specifically, our hypothesis is that the combination of known mathematical models with data-based (DB) models can improve the results of purely MB or DB approaches. First, we investigate the application of a CNN architecture to restore FFDM images, also evaluating the influence of various loss functions and diverse training strategies. Second, we introduce an MBDL approach inspired by a pipeline designed to restore LD mammographic images. We use a variance stabilization transformation (VST) and known system-related parameters to introduce priors implemented as neural network layers. Considering a Poisson-Gaussian noise model, this framework operates within a VST domain, where the noise becomes approximately Gaussian, signal-independent, and with unity variance, enhancing the stability and simplicity of the learning process. Moreover, we propose a bias-residual noise loss function to control the final noise characteristics. Three different CNN architectures were tested and resulted in better performance compared with solely DB approaches. Finally, we also propose an MBDL method to restore mammographic images corrupted with spatially correlated noise. Although further validation is necessary, preliminary results indicate that the MBDL may be suitable for this task. In conclusion, the synergy of MB methods and DB approaches has great potential to be explored within the DL domain, demonstrated by the improved results over models that do not benefit from those priors.

Keywords: Digital breast tomosynthesis. Convolutional neural networks. Deep Learning. Artificial neural networks. Model-based deep learning. Image restoration.

RESUMO

VIMIEIRO, Rodrigo de Barros **Aprendizagem profunda baseada em modelo para restauração de imagens de tomossíntese digital mamária de baixa dose**. 2023. 144p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A tomossíntese digital da mama (DBT) e a mamografia digital de campo total (FFDM) são os exames mais utilizados para rastreamento do câncer de mama. A dose de radiação é uma preocupação, visto que estes sistemas utilizam radiação ionizante. Embora os sistemas atuais operem dentro de margens seguras de radiação, há um desejo crescente de minimizar a dose de radiação sem comprometer a qualidade da imagem. As técnicas de restauração de imagens surgiram como ferramentas para melhorar a qualidade da imagem a partir de aquisições de baixa dose. Os métodos tradicionais de restauração baseiam-se em modelos matemáticos que representam a física de aquisição do sistema. As redes neurais convolucionais (CNN) são capazes de aprender a partir de dados e têm apresentado progresso substancial nos últimos anos. Este trabalho propõe uma estrutura híbrida de aprendizagem profunda baseada em modelo (MBDL) para a restauração de imagens DBT adquiridas com doses reduzidas de radiação, beneficiando-se das vantagens de ambos os campos. Especificamente, nossa hipótese é que a combinação de modelos matemáticos conhecidos com modelos baseados em dados (DB) pode melhorar os resultados de abordagens puramente MB ou DB. Primeiramente, investigamos a aplicação de uma arquitetura CNN para restaurar imagens FFDM, avaliando também a influência de diversas funções de custo e diversas estratégias de treinamento. Em segundo lugar, apresentamos uma abordagem MBDL inspirada em um *framework* projetado para restaurar imagens mamográficas de baixa dose. Usamos um transformada de estabilização de variância (VST) e parâmetros conhecidos relacionados ao sistema para introduzir conhecimentos a *priori* implementados como camadas da rede neural. Três arquiteturas diferentes foram testadas e resultaram em um melhor desempenho em comparação com abordagens exclusivamente baseadas em dados. Finalmente, também propomos um método MBDL para restaurar imagens mamográficas corrompidas com ruído correlacionado espacialmente. Embora seja necessária uma validação adicional, os resultados preliminares indicam que o MBDL pode ser adequado para esta tarefa. Em conclusão, a sinergia entre métodos MB e abordagens de DB tem grande potencial a ser explorada, demonstrado pelos melhores resultados em relação aos modelos que não se beneficiam desses conhecimentos a *priori*.

Palavras-chave: Tomossíntese digital mamária. Redes neurais convolucionais. Aprendizado profundo. Redes neurais artificiais. Aprendizado profundo baseado em modelo. Restauração de imagens.

LIST OF FIGURES

Figure 1 – Architecture of the network used in this study. 46

Figure 2 – Histograms showing the variability of (a) kVp, (b) mAs and (c) breast thickness in the clinical images from the training dataset. 48

Figure 3 – Image of the physical anthropomorphic breast phantom used in this study. The red arrow points to the pieces of calcium oxalate that were placed between the slabs to simulate a cluster of microcalcifications. . . 51

Figure 4 – Illustration of how the anthropomorphic breast phantom dataset was subdivided for the assessment of the MNSE and its decomposition. The metrics were evaluated on all different colored sub-groups of FD, LD and restored images. 52

Figure 5 – Schematic explaining the training process of the models. 54

Figure 6 – Schematic explaining the test procedure for the models. 54

Figure 7 – Illustration of a magnified ROI of a clinical image focusing a small cluster of MC. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv} and (m) FID. (n) model-based restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range. 56

Figure 8 – Illustration of a magnified ROI of a clinical image focusing a small cluster of MC. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv} and (m) FID. (n) model-based restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range. 57

Figure 9 – Illustration of a magnified ROI of an anthropomorphic breast phantom image focusing a simulated cluster of MC. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range. 58

Figure 10 – Illustration of a magnified ROI of an anthropomorphic breast phantom image focusing a simulated cluster of MC. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range. 59

Figure 11 – Illustration of the SNR map of the anthropomorphic breast phantom images. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the maps were adjusted and clipped in the range of 47-120, based on the SNR of the FD image. . . 60

Figure 12 – Illustration of the SNR map of the anthropomorphic breast phantom images. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the maps were adjusted and clipped in the range of 47-120, based on the SNR of the FD image. . . 61

Figure 13 – p-value results among all DL losses, for 75% dose reduction, by paired t-test in terms of (a) MNSE, (b) $\mathcal{R}_{\mathcal{N}}$, and (c) $\mathcal{B}^2$. 64

Figure 14 – p-value results among all DL losses, for 50% dose reduction, by paired t-test in terms of (a) MNSE, (b) $\mathcal{R}_{\mathcal{N}}$, and (c) $\mathcal{B}^2$. 65

Figure 15 – Illustration of the deep neural network architecture utilized in this work. 72

Figure 16 – SNR maps for (a) central DBT projection acquired with the physical anthropomorphic breast phantom at the FD; (b) central projection for the virtual phantom using the OpenVCT software; (c) and (d) are the SNR maps of the same physical and virtual phantoms, respectively, acquired at 50% of the FD. 73

Figure 17 – Normalized noise power spectrum (NNPS) for the physical and virtual PMMA phantoms at the LD and FD. 75

Figure 18 – ROIs extracted from central DBT projections of the physical phantom: (a) LD acquisition; (b) FD acquisition; (c) to (g) DL restoration results for different amounts of training data; (h) restored using the MB pipeline. 76

Figure 19 – Illustration of the losses for the (a) L1 and (b) PL for each training step. Also, SSIM calculated at each training step for both (c) L1 and (d) PL losses. 77

Figure 20 – Illustration of (a) PL, (b) PS loss, (c) total loss and (d) SSIM on the training stage of all the proposed neural networks. 84

Figure 21 – ROIs extracted from the central projection of DBT acquisitions of the physical breast phantom at (a) LD; (b) FD; CNN output with (c) PL3, (d) PL4, (e) PS-1D, (f) PS-2D, (g) the combined loss (PL+PS) and (h) the Model-based (MB) restoration. 85

Figure 22 – Visual representation of the relation between the weighting factor (λ) and the MNAE metric between the PS curves. 86

Figure 23 – Illustration of (a) the mean PS measurement for the DBT central projection acquired with the physical breast phantom at the testing stage for all training realizations and (b) zoom on PS high frequencies (above 1 mm^{-1}) along with the standard deviation (SD). To facilitate the visualization of the SD of each zoomed PS of the DL methods, we plotted them separately from (c) to (g). 87

Figure 24 – Illustration of (a) PS on the physical breast phantom that achieved the lowest MNAE for each restoration method within all training realizations; (b) a zoom on PS high frequencies. 88

Figure 25 – Illustration of the proposed framework. Note that the framework can be applied independently of the network architecture and the illustrated standard denoising CNN is a generic architecture. GAT is the Generalized Anscombe Transformation layer, BS is a layer that normalizes the signal based on fixed values calculated from the entire dataset, Conv is a generic convolutional layer, BN is the commonly known batch normalization, ReLU is the activation function, Add is a generic residual layer, I-GAT the inverse of the GAT and wAdd is a layer similar to the residual layer but with calculated weights. 96

Figure 26 – Scheme illustrating the conventional data-based deep learning framework for image restoration, where the model is meant to learn only from data. Also, the proposed model-based deep learning is separated into two steps. First (1), the model learns in a self-learning framework how to completely remove noise from an image in the VST domain. Here, the target images are simulated using the Noise2Sim framework. Second (2), the pre-trained network is used in the pipeline and can be fine-tuned using the proposed $\mathcal{L}_{\mathcal{BR}}$ loss function to adjust the denoising properties. 97

Figure 27 – $\lambda_{\mathcal{R}}$ tuning plot on the last epoch. The graphic illustrates the network operation points in terms of bias and $\mathcal{R}$ for each $\lambda_{\mathcal{R}}$ value. 105

Figure 28 – Visualization of an ROI for a visual comparison of the DBDL and MBDL restorations for different architectures. Where (a) represents the LD image, (b) the corresponding FD image and (c) the MB method restoration. Figures (d) to (f) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the RED architecture. Figures (g) to (i) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the ResNet architecture. Figures (j) to (l) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the U-Net architecture. Note that each line represents a different architecture. 106

Figure 29 – Illustration of bias and residual noise for each restoration method. The gray dashed lines show points where the MNSE is the same. The green is related to the U-Net architecture, red with the RED and black with the ResNet. 109

Figure 30 – Scheme illustrating how the training dataset was generated. From the 500 phantoms, three sets of projections were generated. The input dataset at 90mAs with independent white noise and two targets with correlated noise. One target at the same dose as the input and the other one at 360mAs. 115

Figure 31 – ROIs of the testing phantom generated with the OpenVCT. Each line is showing a region of interest extracted from: (a) input images at the FD with correlated noise; (b) cGAN #1 restoration at the FD; (c) target FD images with independent noise; (d) cGAN #2 restoration at 4x the FD and (e) target images at 4x the FD with independent noise. 117

Figure 32 – ROIs extracted from DBT central projections of six clinical cases. The first row represents the input images at the standard full dose. The second and third rows represent the output from the networks targeting the independent noise at the same dose (cGAN #1) and four times the dose (cGAN #2), respectively. (a) to (f) illustrate the six different clinical cases. 118

Figure 33 – (a) Power Spectra (PS) calculated at the testing step of the proposed network cGAN #1, where the network performs only noise decorrelation; (b) zoom-in on PS high-frequencies where quantum and electronic noise is known to predominate. 119

Figure 34 – (a) Power Spectra (PS) calculated at the testing step of the proposed network cGAN #2, where the network performs both denoising and decorrelation; (b) zoom-in on PS high-frequencies where quantum and electronic noise is known to predominate 119

Figure 35 – Scheme illustrating the proposed pipeline to generate and insert MC cluster in DBT projections. 125

Figure 36 – The graphic user interface used by readers to perform the two-alternative forced-choice test. One real and one simulated MC cluster are displayed. 126
Figure 37 – (a) Distribution of volume, in number of voxels, of MCs for the entire dataset. (b) Distribution of volume, in number of voxels, of MCs, after the selection criteria. 127
Figure 38 – Illustration of real (first row) and simulated MCs (second row). The ROIs were extracted from DBT slices. 128
Figure 39 – Illustration of 4 slices out of the 15 from the simulated cluster. The figure shows simulated MCs on different depths. 128
Figure 40 – 2-AFC results for each individual reader. The gray region shows the confidence interval for the random selection of a single reader (Binomial distribution with N=100). 129

LIST OF TABLES

Table 1 – Mean SNR values for the anthropomorphic breast phantom images acquired at the LD (for a dose reduction factors of 75% and 50%) and FD. Results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method. . 62

Table 2 – Quantitative analysis of the total MNSE, decomposed into $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$, for the anthropomorphic breast phantom images acquired at the LD (at a dose reduction factor of 75%) and at the FD. Also, the results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method. Confidence interval is shown for p-value=0.05. 62

Table 3 – Quantitative analysis of the total MNSE, decomposed into $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$, for the anthropomorphic breast phantom images acquired at the LD (at a dose reduction factor of 50%) and at the FD. Also, the results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method. Confidence interval is shown for p-value=0.05. 63

Table 4 – Average processing time to run a full restoration on a single raw clinical mammography. 65

Table 5 – SNR values for the real and virtual phantoms. The third column gives the relative errors from the virtual phantom compared to the real acquisition. 74

Table 6 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD on the last epoch. . . 78

Table 7 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on the last epoch. 78

Table 8 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD one on all evaluated epochs. 79

Table 9 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on all evaluated epochs. 79

Table 10 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on the last epoch and were calculated using the new breast mask. 80

Table 11 – MNAE metric between the PS curves (Fig. 24 (b)) of different DL losses and the FD acquisition. Hyper-parameter tuning of the combined loss function weight. 86

Table 12 – Mean value of the total MNSE, $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the best PS in terms of MNAE, described in the last column. The confidence interval is displayed inside the brackets. 88

Table 13 – Mean value of the total MNSE, $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD values, described in the last column. The confidence interval is displayed inside the brackets. 89
Table 14 – Mean value of the total MNSE, $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the lowest $\mathcal{B}^2$, described in the last column. The confidence interval is displayed inside the brackets. 89
Table 15 – MNSE analysis with its decomposition on bias and $\mathcal{R}$ for the DBDL and MBDL approaches and comparison with LD, FD and MB results. Values highlighted mean the lowest bias for each architecture. We also highlighted the closest $\mathcal{R}$ to the value from FD images. 107
Table 16 – SSIM and SNR objective metrics for the DBDL and MBDL approaches and comparison with LD, FD and MB results. We highlighted the closest SSIM and SNR to the FD value. 109
Table 17 – Training time, in hours, for DBDL, MBDL and MBDL$^{\lambda}$ modes. Inference time for all frameworks, in seconds 110
Table 18 – Mean value of the MNSE calculated for all DBT projections of the synthetic breast phantom used for testing. The confidence interval is displayed inside the brackets. Input means the original correlated 90mAs image, output #1 the result cGan for 90mAs images, output #2 the cGan result for 360mAs images, target #1 the white noise 90mAs image and target #2 the white noise 360mAs image. 120

LIST OF ABBREVIATIONS AND ACRONYMS

AI	artificial intelligence
CT	computed tomography
CV	computer vision
CNN	convolutional neural network
CAD	computer-aided diagnosis
cycle-GAN	cycle generative adversarial network
DBT	digital breast tomosynthesis
DL	deep learning
DB	data-based
DNN	deep neural network
DBDL	data-based deep learning
FFDM	full-field digital mammography
FD	full-dose
FID	Fréchet inception distance
GPU	graphics processing unit
GT	ground-truth
GAT	generalized Anscombe transformation
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
LD	low-dose
LLM	large language models
ML	machine learning
MRI	magnetic resonance imaging
MC	microcalcifications
MSE	mean squared error
MNSE	mean normalized squared error
MAE	mean absolute error
MB	model-based
MBIR	model-based iterative reconstruction

MBDL	model-based deep learning
NN	neural network
PL	perceptual loss
PS	power spectrum
ResNet	residual network
RED	residual encoder-decoder
RN	residual noise
SNR	signal-to-noise ratio
SSIM	structural similarity index measure
VCT	virtual clinical trials
VST	variance stabilization transformation
2D	two-dimensional
3D	three-dimensional
2-AFC	two-alternative forced-choice

CONTENTS

1 INTRODUCTION . . . 31

1.1 Motivation . . . 31

1.2 Objectives . . . 33

1.2.1 General objectives . . . 33

1.2.2 Specific objectives . . . 33

1.3 Main contributions . . . 34

1.4 Thesis organization . . . 38

1.5 Other contributions . . . 39

2 PAPER 1: IMPACT OF LOSS FUNCTIONS ON THE PERFORMANCE OF A DEEP NEURAL NETWORK DESIGNED TO RESTORE LOW-DOSE DIGITAL MAMMOGRAPHY . . . 41

2.1 Abstract . . . 41

2.2 Introduction . . . 42

2.3 Theoretical background . . . 45

2.3.1 Degradation model . . . 45

2.3.2 Restoration model . . . 46

2.3.2.1 Model-based . . . 46

2.3.2.2 Data-based . . . 46

2.3.3 Loss functions . . . 47

2.3.3.1 Mean squared error . . . 47

2.3.3.2 Mean absolute error or ℓ_1 . . . 48

2.3.3.3 Structural similarity index . . . 48

2.3.3.4 Perceptual loss . . . 48

2.3.3.5 Adversarial loss . . . 49

2.3.3.6 Fréchet inception distance . . . 49

2.4 Materials & methods . . . 49

2.4.1 Training dataset . . . 50

2.4.2 Testing dataset . . . 50

2.4.3 Figures of merit . . . 51

2.4.4 Implementation details . . . 53

2.5 Results & discussions . . . 55

2.5.1 Visual analysis . . . 55

2.5.2 Quantitative evaluation . . . 59

2.6 Conclusion . . . 66

3 PAPER 2: ASSESSMENT OF TRAINING STRATEGIES FOR CONVOLUTIONAL NEURAL NETWORK TO RESTORE LOW-DOSE DIGITAL BREAST TOMOSYNTHESIS PROJECTIONS 69
3.1 Abstract . . . 69
3.2 Introduction . . . 69
3.3 Materials & methods . . . 70
3.3.1 Training and testing datasets . . . 70
3.3.1.1 Training . . . 71
3.3.1.2 Testing . . . 72
3.3.2 Network and implementation details . . . 72
3.3.3 Objective metrics . . . 74
3.4 Results & discussions . . . 74
3.4.1 Virtual phantom validation . . . 74
3.4.2 Visual analysis . . . 75
3.4.3 Quantitative evaluation . . . 76
3.5 Conclusions . . . 79

4 PAPER 3: IMPOSING NOISE CORRELATION FIDELITY ON DIGITAL BREAST TOMOSYNTHESIS RESTORATION THROUGH DEEP LEARNING TECHNIQUES . . . 81
4.1 Abstract . . . 81
4.2 Introduction . . . 81
4.3 Materials & methods . . . 82
4.3.1 Training and testing datasets . . . 82
4.3.1.1 Training . . . 82
4.3.1.2 Testing . . . 82
4.3.2 Network and implementation details . . . 83
4.4 Results & discussions . . . 83
4.5 Conclusions . . . 90

5 PAPER 4: LEARNING IN A VARIANCE STABILIZATION DOMAIN TO RESTORE LOW-DOSE DIGITAL BREAST TOMOSYNTHESIS PROJECTIONS . . . 91
5.1 Abstract . . . 91
5.2 Introduction . . . 91
5.3 Theoretical background . . . 94
5.4 Materials & methods . . . 96
5.4.1 Architecture . . . 96
5.4.2 Training . . . 97
5.4.2.1 Loss functions . . . 98

5.4.3 Dataset . . . 99
5.4.3.1 Patient FD/LD pairs . . . 99
5.4.3.2 Synthetic FD/LD pairs . . . 99
5.4.3.3 Phantom FD/LD pairs . . . 100
5.4.3.4 Uniform phantoms . . . 100
5.4.4 Experimental setup . . . 101
5.4.4.1 MBDL . . . 101
5.4.4.2 DBDL . . . 102
5.4.4.3 MB . . . 102
5.4.5 Figures of merit . . . 102
5.5 Results & discussions . . . 103
5.5.1 $\lambda_{\mathcal{R}}$ Optimization . . . 104
5.5.2 Visual comparison . . . 104
5.5.3 Quantitative analysis . . . 107
5.6 Conclusion . . . 110

6 PAPER 5: SUPPRESSING NOISE CORRELATION IN DIGITAL BREAST TOMOSYNTHESIS USING CONVOLUTIONAL NEURAL NETWORK AND VIRTUAL CLINICAL TRIALS . . . 113
6.1 Abstract . . . 113
6.2 Introduction . . . 113
6.3 Materials & methods . . . 114
6.3.1 Training and testing datasets . . . 114
6.3.1.1 Training . . . 114
6.3.1.2 Testing . . . 115
6.3.2 Network and implementation details . . . 116
6.4 Results & discussions . . . 116
6.5 Conclusions . . . 120

7 PAPER 6: COMPUTATIONAL METHOD TO ARTIFICIALLY INSERT CLUSTERS OF MICROCALCIFICATIONS IN DIGITAL BREAST TOMOSYNTHESIS . . . 123
7.1 Abstract . . . 123
7.2 Introduction . . . 123
7.3 Materials & methods . . . 124
7.3.1 Simulation pipeline . . . 124
7.3.2 Two-alternative forced-choice . . . 125
7.3.3 Lesion-free and lesion-present dataset . . . 126
7.3.4 Microcalcification dataset . . . 126
7.3.5 Image reconstruction . . . 127

7.4 Results & discussions . . . 127

7.4.1 Simulated images . . . 127

7.4.2 Two-alternative forced-choice study . . . 128

7.5 Conclusions . . . 129

8 CONCLUSIONS . . . 131

REFERENCES . . . 133

1 INTRODUCTION

1.1 Motivation

Artificial intelligence (AI), and more specifically the machine learning (ML) area, has been substantially improved in the field of medicine (TOPOL, 2019; RAJKOMAR; DEAN; KOHANE, 2019), mainly in medical imaging (LEE; FUJITA, 2020). This advance for medical imaging applications is explained due to the excellent results of these techniques in computer vision (CV) tasks not related to health, such as in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Even more recently, the usage of large language models (LLM) has revolutionized the general area of AI (RADFORD *et al.*, 2018; RADFORD *et al.*, 2019; BROWN *et al.*, 2020; DEVLIN *et al.*, 2018).

With the advent of deep learning (DL) techniques, which have been replacing conventional CV techniques, numerous applications are being developed for diagnostic imaging systems, such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, mammography, etc. These applications cover areas such as medical image classification (CIOMPI *et al.*, 2017; SHEN *et al.*, 2019a; MCKINNEY *et al.*, 2020), computer-aided diagnosis (CAD) (HUA *et al.*, 2015; AL-MASNI *et al.*, 2018; CHAN; SAMALA; HADJIISKI, 2019), tomographic image reconstruction (WU *et al.*, 2017; ADLER; ÖKTEM, 2018), image segmentation (RONNEBERGER; FISCHER; BROX, 2015), noise filtering (RAN *et al.*, 2019), artifact reduction (ZHANG; YU, 2018), radiation dose reduction (CHEN *et al.*, 2017a; YANG *et al.*, 2018a), among many others (LANGLOTZ *et al.*, 2019).

In the CT field, a large number of works have been published in the area of DL exploring the recent capabilities of convolutional neural network (CNN) (CHEN *et al.*, 2017a; CHEN *et al.*, 2017b; SHAN *et al.*, 2018; YANG *et al.*, 2018a; SHAN *et al.*, 2019; MAN *et al.*, 2019; YOU *et al.*, 2019; WU *et al.*, 2021). Similarly, in mammography systems, these tools are also being applied (SAMALA *et al.*, 2018; SHEN *et al.*, 2019b; CHAN; SAMALA; HADJIISKI, 2019; HARVEY *et al.*, 2019; AREFAN *et al.*, 2020).

Breast cancer kills approximately 685,000 women annually around the world and this disease is the leading cause of cancer-related deaths for women (WHO, 2023). Screening procedures aim to examine a certain asymptomatic population to detect cancer at an early stage. According to the WHO (2023), the five-year survival rate[1] is 80% when screening procedures and basic treatments are taken.

[1] "The percentage of people in a study or treatment group who are alive five years after they were diagnosed with or started treatment for a disease, such as cancer. The disease may or may not have come back." Source: National Cancer Institute Dictionaries.

Full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT) are the preferred exam for breast cancer screening (SKAANE *et al.*, 2013; FRIEDEWALD *et al.*, 2014). While the former generates a two-dimensional (2D) projection image of the breast, the latter produces several projections at different angles and uses a reconstruction algorithm to obtain slices of a three-dimensional (3D) volume of the breast. (VEDANTHAM *et al.*, 2015; HOOLEY; DURAND; PHILPOTTS, 2017). Both systems use ionizing radiation at approximately the same radiation dose. Historically, several applications have been applied to these systems to improve image quality and consequently the diagnosis precision. Moreover, both require a compromise to keep the radiation dose at lower levels while keeping satisfactory image quality.

Among the various applications of deep neural networks (DNNs), from image generation to feature extraction, image restoration and tomographic reconstruction are extremely important (WANG *et al.*, 2018; RAVISHANKAR; YE; FESSLER, 2019; WANG; YE; MAN, 2020). In the context of medical image restoration, these problems have been solved by traditional methods like iterative reconstruction (DAS *et al.*, 2010; XU *et al.*, 2015; ZHENG; FESSLER; CHAN, 2017) and denoising techniques (WU; MAINPRIZE; YAFFE, 2012a; BORGES *et al.*, 2017; BORGES *et al.*, 2018). However, data-based (DB) methods have been successfully applied in the field achieving excellent results (WU *et al.*, 2017; KANG; MIN; YE, 2017; CHEN *et al.*, 2017a; WOLTERINK *et al.*, 2017; CHEN *et al.*, 2017b; KANG *et al.*, 2018; SHAN *et al.*, 2018; SHAN *et al.*, 2019; YIN *et al.*, 2019; WU *et al.*, 2021).

The motivation of image processing techniques in both CT and DBT systems, in general, is related to the improvement of image quality for a better diagnosis or to mitigate patient health risks such as in applications to reduce radiation dose. Several works based on domain-specific mathematical models have been proposed to achieve these purposes in areas such as model-based iterative reconstruction (MBIR) methods (LEVAKHINA *et al.*, 2013; XU *et al.*, 2015; ZHENG; FESSLER; CHAN, 2017; GARRETT *et al.*, 2018), radiation dose reduction techniques (WANG *et al.*, 2006; MANDUCA *et al.*, 2009; BORGES *et al.*, 2018) and so on. DL techniques have emerged in the image processing area because they can learn different tasks directly from the data in a supervised or unsupervised manner. Although there are neural networks (NNs) that were applied several years ago, it was not until recently that they achieved promising results and became popular. This overwhelming interest is partially explained due to the increase in computational power, allowing deeper networks and also the greater availability of data. Although recently we have more access to data, in the medical imaging scenario, data availability is still a concern (WANG; YE; MAN, 2020).

More specifically in areas with DB applications, Liu *et al.* (2018), Sahu *et al.* (2019) and Gao, Fessler and Chan (2021) proposed DL methods to restore low-dose (LD)

DBT images. These works mainly focused on the projection domain, *i.e.*, applying those techniques to restore projection images. Similarly, Green *et al.* (2019) proposed a CNN to restore LD FFDM images. Although they reached very promising results for DBT compared to traditional methods, some limitations include the fact that the networks were trained with breast specimens or physical/simulated phantoms. Different from the other, for FFDM, Green *et al.* (2019) adapted a noise injection technique from digital chest radiography. Although the restoration process in these works depends mainly on the provided data for network learning, incorporating prior knowledge into the network's architecture could potentially enhance its performance.

This field is known as model-based deep learning (MBDL), where according to Shlezinger *et al.* (2023), this area takes advantage of known mathematical models and DB methods. Several works have been investigating this strategy (WU *et al.*, 2017; KANG *et al.*, 2018; CHEN *et al.*, 2018; WÜRFL *et al.*, 2018; ADLER; ÖKTEM, 2018; GONG *et al.*, 2018; ZHANG *et al.*, 2019; WU *et al.*, 2021). Xia *et al.* (XIA *et al.*, 2023) recently have done an extensive review of MBDL for LD CT, comparing methods and demonstrating the advantages of both DB and model-based (MB) approaches.

To train DL architectures for some specific task, like image restoration, usually it is necessary to have access to a large amount of data. As previously mentioned, the medical imaging field has limitations in terms of data availability. Also, data diversity is expected so that the DB models can generalize well for different cohorts. Since MBDL accounts with specific domain knowledge, due to the availability of precise mathematical formulations proposed for specific tasks, the learning space is reduced and then the ML algorithms can learn the data representation with less data (SHLEZINGER *et al.*, 2023). Moreover, pure CNNs are known as black boxes, making their application to the clinical routine more difficult. The combination of DB models and physics/MB methods takes advantage of both approaches enabling MBDL methods to be safer, more robust, and more interpretable (XIA *et al.*, 2023).

1.2 Objectives

1.2.1 General objectives

The objective of this work is to propose a framework to restore LD DBT images using an MBDL approach. This framework is composed of well-known DB methods, *e.g.*, a CNN, and takes advantage of popular mathematical formulations for image restoration.

The specific objectives of this thesis are:

1.2.2 Specific objectives

1. Create a dataset of mammography images to train a DB model for image restoration;

2. Investigate the influence of different loss functions in the task of restoration of LD digital mammography;
3. Investigate the influence of dataset size, cross-validation and early stop for DBT image restoration;
4. Demonstrate the importance of noise modeling in simulated images for good image restoration using DB models;
5. Investigate the image restoration results in the frequency domain;
6. Propose image restoration techniques to reduce noise spatial correlation through DL models;
7. Propose a MBDL framework to restore LD DBT images;
8. Compare the proposed framework with solely DB methods;
9. Propose a framework to restore LD DBT images corrupted with spatially correlated noise;
10. Create a method to artificially insert clusters of microcalcifications (MC) in DBT images;

1.3 Main contributions

In this section, I correlate the previously stated general and specific objectives with the main contributions of this work. A general overview of each published manuscript is made and it shows the respective association with the objectives.

Since DL methods are strongly dependent on the network architecture, loss function and training strategies, in the first part of this thesis we used a commonly known residual network (ResNet) to investigate the impact of several loss functions for the restoration of FFDM images acquired with reduced radiation dose. In this work, **Paper 1**, we proposed a DL framework to restore LD FFDM images which are meant to learn solely from the provided data. Loss functions for image restoration have been previously investigated for natural images but to the best of our knowledge not for medical images in the field of mammography.

As the first step, we built a dataset with pairs of LD and full-dose (FD) images. We used an in-house method to simulate the LD acquisitions in 400 retrospective clinical images. Such dataset contains a great diversity of radiographic factors such as kVp and mAs and different breast thicknesses. This is important for ML algorithms in the sense of generalization capacity.

A total of 11 losses were investigated in terms of two objective metrics. First, we used the signal-to-noise ratio (SNR) and second the decomposition of mean normalized squared error (MNSE) in bias and residual noise (RN). Such decomposition was able to show that image-fidelity losses such as mean squared error (MSE) and mean absolute error (MAE) had strong denoising properties, decreasing the noise at the cost of increasing the bias. On the other hand, it was shown that visual content losses such as perceptual loss (PL), adversarial, Fréchet inception distance (FID) and structural similarity index measure (SSIM) had a good trade-off in terms of RN matching with the FD, keeping low levels of bias.

This first part was an important step to show the importance of loss functions for DL methods so when we compare standard data-based deep learning (DBDL) approaches, we can choose a relevant method in this framework. This work was published by *Artificial Intelligence in Medicine* journal. This journal is internationally recognized (Impact Factor: 7.5) and is classified as A1 by CAPES Qualis rank. Chapter 2 shows the complete work that was published. Code is available on GitHub.

Since we have limited access to clinical data, some important training designs are difficult to investigate. In **Paper 2**, we investigated the impact of dataset size, early stop and cross-validation on the restoration of DBT projections. We used virtual clinical trials (VCT) software to simulate 500 breast phantoms, yielding a total of 1 million patches used for training. Even though we used simulated data for training, a physical anthropomorphic breast phantom was used to measure the denoising properties through the MNSE decomposition in bias and RN. We also showed in this work the importance of accurate noise modeling for virtual images when DB models are used.

Through objective metrics, we observed that training the model several times leads to distinct results. Depending on the task, the model can match the RN of the target or achieve low bias values. Early stops are feasible, especially in cases where low bias values are desired. Also, the largest dataset, with 1 million patches, did not improve the results overall, and the model was able to get good results with approximately a quarter of this amount.

All those findings were extremely important for further comparison between DBDL and MBDL methods. Since the ML field has many hyper-parameters, investigation and an extensive search for the best configuration are important. Then, we can perform a fair comparison between DBDL and MBDL methods, using the correct loss function, dataset size, and training strategy in general. Chapter 3 shows the complete work that was published in the conference proceedings. This paper was published and presented at the SPIE Medical Imaging conference, in 2023. Code is available on GitHub.

In Paper 1 and Paper 2, we observed some artifacts related to the PL. Those artifacts resemble a chessboard pattern and were recently stated by Xia *et al.* (2023)

as well. This artifact is a consequence of the noise spatial correlation that the CNN generates in the convolution process and can be observed in the frequency domain. Such fact motivated the development of **Paper 3**, which proposed a method to restore LD DBT projections and imposes noise properties fidelity in the frequency domain between the original FD and restored images. The method consists of a power spectrum (PS) loss combined with the PL. We built a dataset, following the same procedure stated in Paper 1, with pairs of LD and FD images. We found that the combination of both losses yields good results in terms of PS fidelity with respect to the FD target image, also achieving good results of RN and bias. This paper was published and presented at the 16th International Workshop on Breast Imaging 2022. Chapter 4 shows the complete work that was published in the conference proceedings. Code is available on GitHub.

In the MBDL scenario, **Paper 4** proposes a new approach to restore LD DBT projections. We hypothesize that priors added to the network, in terms of layers, improve the stability and final results. In this framework, we use precise mathematical formulations developed over the years in combination with recent DB approaches such as neural networks, as presented in the previous works. With the best of the two approaches, *i.e.*, the specific domain knowledge and stability for the MB approaches, and the great learning capacity from DB models, the combination of both yields excellent results.

Since the proposed framework is agnostic to network architecture, we used three different neural network architectures to perform a comparison between the DB and MB approaches. We used the commonly known ResNet, residual encoder-decoder (RED) and U-Net. For all architectures, we introduced new specialized layers to perform the variance stabilization through the generalized Anscombe transformation (GAT), the inverse GAT, and a residual layer with calculated weights. All those layers were inspired by a MB pipeline proposed by Borges *et al.* (2017).

In general, there are two steps in the restoration framework. The first one, the networks are trained with data in the variance stabilization transformation (VST) domain in a self-learning fashion. In this domain, the noise becomes approximately Gaussian, signal-independent, and with unity variance, making the NN learning procedure much easier compared with conventional methods. This is done since we do not have noise-free images to use as ground-truth (GT) and the MB pipeline was designed in such a way that the filtering method completely removes the noise of the image. Second, the networks are fine-tuned end-to-end with a novel bias-RN loss function, proposed in this work, to adjust the trade-off between noise suppression and signal blur.

For each step, we used a specific dataset. First, in the self-learning mode, we used a clinical dataset with simulated LD images. For the fine-tuning, we used synthetic images from a VCT software, similar to Paper 2. Lastly, we used a real anthropomorphic phantom to evaluate all the proposed models, similar to Paper 1.

For comparison, we used the DBDL approach similar to what we presented in Paper 1, using a combination of a MAE and PL, since it achieved the best results. For the training strategy, we followed Paper 2, where we trained the DB approach for 60 epochs for each loss, since we are targeting a RN match between the restored image and the FD.

We also used the MNSE decomposition, as in the other papers, and the proposed MBDL approach had superior results when compared to models trained solely with data without any priors. Still in objective metrics, the hybrid approach also achieved superior results in terms of SSIM and SNR. The new framework was also superior in terms of training time, graphics processing unit (GPU) memory utilization and it was demonstrated to be more stable. The stability of MBDL was visualized when different architectures produced objective metrics very similar to each other. This contrasts with what we stated in Paper 2 for DBDL methods that achieved different results for different realizations even for the same architecture.

This work was submitted to *IEEE Transactions on Medical Imaging* journal and it is currently under review. Chapter 5 shows the complete work that was submitted. Code is available on GitHub.

As stated in Paper 3, noise spatial correlation has to be considered in mammography systems. Especially for indirect detectors that are affected by pixel cross-talk. Different from the previous papers, which were proposed on a system with direct x-ray detection, known to have white-noise properties, in **Paper 5**, we proposed a framework to restore DBT projections that were degraded by frequency-dependent (correlated) noise. We also consider this work as part of an MBDL field. With the mathematical models that estimate and simulate noise correlation on those images, we can train a network to restore images that were degraded in this scenario. Training with data acquired only in the equipment itself would not be sufficient for such a task.

Using a cycle generative adversarial network (cycle-GAN) framework, we trained two networks to perform a domain transform between correlated and independent images. We followed all the mathematical models developed in the work of Borges *et al.* (2019) to estimate and simulate the noise correlation kernel so we could apply it to the simulated images on the VCT software. Since we have the target PS curves, this work shows that the CNN was able to perform both denoising and decorrelation with noise frequency components very close to the target image. We also illustrated some results for clinical images, showing the potential of the framework, however, further clinical studies must be performed to evaluate the real impact of the restoration on object localization and detection. This paper was published and presented at the 16th International Workshop on Breast Imaging 2022. Chapter 6 shows the complete work that was published in the conference proceedings. Code is available on GitHub.

As discussed in Paper 5, it is necessary to clinically evaluate the proposed restoration through human perception studies. In **Paper 6**, we presented a method to artificially insert MC clusters in normal exams of DBT. Such a tool is important as a first step in the human perception study since only a small fraction of clinical images have positive cases and the exact location of the lesion is not always known.

In this method, the user can control the size contrast and location of all MCs. The MC dataset was obtained by segmenting the 3D structures from clinical cases acquired in a prone stereotactic breast biopsy system. We used our in-house reconstruction toolbox[2] to both project and back-project the data. A two-alternative forced-choice (2-AFC) study[3] showed a success rate was 53.8%, suggesting that readers, on average, could not distinguish between real and simulated MCs. Such work is extremely important since all the previous works have as one of the objectives improving cancer detection or not decreasing it. This paper was published and presented at the SPIE Medical Imaging conference, in 2023. Chapter 7 shows the complete work that was published in the conference proceedings. Code is available on GitHub.

1.4 Thesis organization

This document is organized in 8 chapters. Chapters 2 to 7 present the published or submitted scientific publications. The selected manuscripts were:

1. Hongming Shan†, **Rodrigo de Barros Vimieiro**†, Lucas Rodrigues Borges, Marcelo Andrade da Costa Vieira, and Ge Wang. "Impact of loss functions on the performance of a deep neural network designed to restore low-dose digital mammography." Artificial Intelligence in Medicine, vol. 142, p. 102555, 2023. Available at: https://doi.org/10.1016/j.artmed.2023.102555.

2. **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Ge Wang, and Marcelo Andrade da Costa Vieira. "Assessment of training strategies for convolutional neural network to restore low-dose digital breast tomosynthesis projections." In Medical Imaging 2022: Physics of Medical Imaging, vol. 12031, p. 470-479. SPIE, 2022. Available at: https://doi.org/10.1117/12.2609 945.

3. **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Ge Wang, and Marcelo Andrade da Costa Vieira. "Imposing noise correlation fidelity on digital breast tomosynthesis restoration through deep learning techniques." In 16th International Workshop on Breast Imaging (IWBI2022), vol. 122861, p. 13. SPIE, 2022. Available at: https://doi.org/10.1117/12.2626634.

[2] www.github.com/LAVI-USP/pyDBT

[3] www.github.com/LAVI-USP/2AFC_Interface

†Shared first authorship.

4. **Rodrigo de Barros Vimieiro**, Chuang Niu, Hongming Shan, Lucas Rodrigues Borges, Ge Wang, and Marcelo Andrade da Costa Vieira. "Learning in a variance stabilization domain to restore low-dose digital breast tomosynthesis projections." (Submitted to *IEEE Transactions on Medical Imaging* journal and it is currently under review).

5. **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Renato França Caron, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Ge Wang, and Marcelo Andrade da Costa Vieira. "Suppressing noise correlation in digital breast tomosynthesis using convolutional neural network and virtual clinical trials." In 16th International Workshop on Breast Imaging (IWBI2022), vol. 12286, p. 314-320. SPIE, 2022. Available at: https://doi.org/10.1117/12.2625357.

6. **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Renato França Caron, Silvia Maria Prioli De Souza Sabino, Ge Wang, and Marcelo Andrade da Costa Vieira. "Computational method to artificially insert clusters of microcalcifications in digital breast tomosynthesis." In Medical Imaging 2023: Physics of Medical Imaging, vol. 12463, p. 572-578. SPIE, 2023. Available at: https://doi.org/10.1117/12.2653857.

Chapter 8 offers general and specific conclusions. In this chapter, I also present the proposed future work.

1.5 Other contributions

During the period of this work, other contributions, beyond the main ones previously stated, were also proposed. The manuscripts are:

1. Arthur Chaves da Costa, **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Bruno Barufaldi, Andrew Douglas Arnold Maidment, and Marcelo Andrade da Costa Vieira. "Assessment of video frame interpolation network to generate digital breast tomosynthesis projections." In 16th International Workshop on Breast Imaging (IWBI2022), vol. 12286, p. 329-336. SPIE, 2022. Available at: https://doi.org/10.1117/12.2625748.

2. Denis Henrique Pinheiro Salvadeo, **Rodrigo de Barros Vimieiro**, Marcelo Andrade da Costa Vieira, and Andrew Douglas Arnold Maidment. "Bayesian reconstruction for digital breast tomosynthesis using a non-local Gaussian Markov random field a priori model." In Medical Imaging 2019: Physics of Medical Imaging, vol. 10948, p. 1246-1251. SPIE, 2019. Available at: https://doi.org/10.1117/12.2513140.

3. Renann de Faria Brandão, **Rodrigo de Barros Vimieiro**, Lucas Rodrigues Borges, Renato França Caron, and Marcelo Andrade da Costa Vieira. “Evaluating the simulation of radiation dose reduction in a digital breast tomosynthesis system featuring an amorphous silicon (a-Si) detector.” Revista Brasileira de Física Médica, vol. 13, no. 2, p. 30-34, 2019. Available at: https://doi.org/10.29384/rbfm.2019.v13.n2.p30-34.

2 PAPER 1: IMPACT OF LOSS FUNCTIONS ON THE PERFORMANCE OF A DEEP NEURAL NETWORK DESIGNED TO RESTORE LOW-DOSE DIGITAL MAMMOGRAPHY

The material presented in this chapter was published in: Shan, Hongming[†], Rodrigo de Barros Vimieiro[†], Lucas Rodrigues Borges, Marcelo Andrade da Costa Vieira, and Ge Wang. “Impact of loss functions on the performance of a deep neural network designed to restore low-dose digital mammography.” Artificial Intelligence in Medicine, vol. 142, p. 102555, 2023. Available at: https://doi.org/10.1016/j.artmed.2023.102555. The author obtained permission for total and/or partial reproduction of the content from the publisher Elsevier.

2.1 Abstract

Digital mammography is currently the most common imaging tool for breast cancer screening. Although the benefits of using digital mammography for cancer screening outweigh the risks associated with the x-ray exposure, the radiation dose must be kept as low as possible while maintaining the diagnostic utility of the generated images, thus minimizing patient risks. Many studies investigated the feasibility of dose reduction by restoring low-dose images using deep neural networks. In these cases, choosing the appropriate training database and loss function is crucial and impacts the quality of the results. In this work, we used a standard residual network (ResNet) to restore low-dose digital mammography images and evaluated the performance of several loss functions. For training purposes, we extracted 256,000 image patches from a dataset of 400 images of retrospective clinical mammography exams, where dose reduction factors of 75% and 50% were simulated to generate low and standard-dose pairs. We validated the network in a real scenario by using a physical anthropomorphic breast phantom to acquire real low-dose and standard full-dose images in a commercially available mammography system, which were then processed through our trained model. We benchmarked our results against an analytical restoration model for low-dose digital mammography. Objective assessment was performed through the signal-to-noise ratio (SNR) and the mean normalized squared error (MNSE), decomposed into residual noise and bias. Statistical tests revealed that the use of the perceptual loss (PL4) resulted in statistically significant differences when compared to all other loss functions. Additionally, images restored using the PL4 achieved the closest residual noise to the standard dose. On the other hand, perceptual loss PL3, structural similarity index (SSIM) and one of the adversarial losses achieved the lowest bias for both dose reduction factors. The source code of our deep neural network is available at https://github.com/WANG-AXIS/LdDMDenoising.

[†]Shared first authorship.

2.2 Introduction

Early diagnosis of breast cancer is crucial to improve the survival rate. The expansion of breast screening programs contributed to improve this rate, which has been significantly increased in the last years (SAADATMAND *et al.*, 2015). This disease is still the main cause of cancer-related deaths among women and screening mammography is the primary tool for detecting tumors at early stages, especially for women at the age of 50-69 (WHO, 2023).

Full-field digital mammography and digital breast tomosynthesis are the most common imaging tools for breast cancer screening (MICHELL; BATOHI, 2018). In these systems a small dose of x-ray radiation is used to generate projections of the breast, which are then interpreted by a radiologist (VEDANTHAM *et al.*, 2015). Although the radiation dose levels are kept within a safe margin, it is desirable to keep the dose as low as possible while maintaining satisfied image quality to fulfill the clinical screening purposes (IAEA, 2018). However, reducing radiation doses can degrade image quality, limiting the performance of the radiologist on searching and characterizing subtle lesions (HAUS; YAFFE, 2000; HUDA *et al.*, 2003; JR *et al.*, 2007; CHAN *et al.*, 2020).

Several works in the field of medical imaging investigate the potential of reducing radiation dose by restoring low-dose (LD) exams to achieve image quality comparable to the ones at the clinical routine. Some proposals, in the field of computed tomography (KALRA *et al.*, 2003; MANDUCA *et al.*, 2009; LI *et al.*, 2014) and digital mammography (WU; MAINPRIZE; YAFFE, 2012a; BORGES *et al.*, 2017; BORGES *et al.*, 2018), evaluated this technique in a model-based (MB) approach, using restoration methods through denoising techniques to improve image quality. In Borges *et al.* (2017) and Borges *et al.* (2018), the authors proposed a pipeline to restore LD mammography through a variance stabilizing transformation (VST). Recently, it was shown that the denoising method may improve the localization of microcalcifications (MC) in these exams (BORGES *et al.*, 2020).

With the rapid development of deep learning techniques, in particular the convolutional neural networks (CNNs), many studies have proposed algorithms to improve the quality of LD images and achieved comparable, or even better results, than the MB ones (WU *et al.*, 2017; KANG; MIN; YE, 2017; CHEN *et al.*, 2017a; WOLTERINK *et al.*, 2017; CHEN *et al.*, 2017b; YANG *et al.*, 2018a; KANG *et al.*, 2018; SHAN *et al.*, 2018; YANG *et al.*, 2018b; SHAN *et al.*, 2019; YIN *et al.*, 2019). This new data-based approach takes advantage of learning features directly from a dataset rather than explicitly applying advanced feature extraction techniques or modeling the system mathematically. One important constraint of data-based techniques is the necessity of a great number of images and a diverse dataset for the training process (SUN *et al.*, 2017). In the field of medical imaging, access to large datasets is limited and some techniques, such as data augmentation and transfer learning, have been applied to increase the size of the dataset (LEE; FUJITA, 2020; COSTA *et al.*, 2020). Moreover, when it comes to LD image restoration using supervised learning, acquisitions at different dose levels are required in the training process. Although it may be possible to create experimental LD/FD imaging protocols, such as in Wolterink *et al.* (2017), exposing the patient multiple times increases the risks of induced cancer.

Thus, in the field of computed tomography, a common approach is to train these deep networks using clinical data (MCCOLLOUGH *et al.*, 2017), where LD images are obtained by injecting extra noise into the standard full-dose (FD) projections (SHAN *et al.*, 2019). In the field of mammography, it is common to use breast specimen Liu *et al.* (2018), physical phantoms Gao, Fessler and Chan (2021) or virtual clinical trials (VCT) software (SAHU *et al.*, 2019) to generate low-dose and full-dose image pairs for the training process of these deep networks. In Green *et al.* (2019) the authors adapted a noise injection technique from digital chest radiography to simulate ultra-low-dose mammography acquisitions and thus trained a CNN for denoising.

When it comes to the deep neural network (DNN) structure and the training step for the restoration process, there are two key components: the network architecture and loss function. The former determines the complexity of the denoising model, while the latter controls the learning process. Thus, the loss function has a direct impact on image quality and it is relatively more important than the network architecture for the task of image restoration (SHAN *et al.*, 2018).

As the restoration of LD mammography is composed of a denoising process, most image translation models can be adapted to this task, such as residual encoder-decoder convolutional neural networks (RED-CNN) (CHEN *et al.*, 2017a), U-net (RONNEBERGER; FISCHER; BROX, 2015) and dense networks (HUANG *et al.*, 2017). Even though the denoising process is part of the restoration process, the main goal of such a task is to map LD images to standard FD. That is where the loss function plays an important role in measuring the similarity between the image pair. Commonly-used loss functions include error visibility methods, such as mean squared error (MSE) and mean absolute error (MAE); structural similarity methods, such as structural similarity index (SSIM) (WANG *et al.*, 2004) and complex Wavelet SSIM index (CW-SSIM) (WANG; SIMONCELLI, 2005); information-theoretical methods, such as IFC (SHEIKH; BOVIK; VECIANA, 2005) and VIF Sheikh and Bovik (2006); and DNN methods, such as perceptual loss (PL) (JOHNSON; ALAHI; FEI-FEI, 2016), adversarial loss (GOODFELLOW *et al.*, 2014), Fréchet inception distance (FID) (MATHIASEN; HVILSHØJ, 2020), and image quality transformer (IQT) (CHEON *et al.*, 2021). In (ZHAO *et al.*, 2016), the authors investigated commonly-used losses for image restoration with neural networks for natural images. In (DING *et al.*, 2021), the authors also investigated several image quality assessment methods as loss function for low-level computer vision tasks on natural images.

However, in contrast with natural images, image restoration for medical images, more specifically for mammography, is a meticulous task, where subtle structures such as MC are extremely important and must be preserved in the restoration process. Moreover, it is desired that the noise properties of the restored image match the FD ones, which is also important for radiologists in the clinical routine. This was demonstrated in (NAGARE *et al.*, 2021), where the authors proposed a loss function that takes into account the bias and noise variance for the restoration algorithm.

The objective of this work is to investigate different loss functions of the DNN and their impact on the performance of the restoration of LD mammography images. We assess the efficiency of each loss function measuring the signal-to-noise ratio (SNR) and the mean normalized squared error (MNSE), decomposed into residual noise and bias, after the restoration process. To this end, we used a commonly known residual convolutional neural network (ResNet) He *et al.* (2016), illustrated on Fig. 1. The reasons why we used this model are two-fold: (1) it has been extensively used for image restoration, and (2) we could avoid any potential bias of a new network for evaluating the impact of loss functions. For the learning process, we build a clinical dataset from retrospective mammography exams, combined with a computational method to simulate LD acquisitions pairs (BORGES *et al.*, 2016; BORGES *et al.*, 2017). The adopted method performs noise injection in a variance-stabilizing domain, avoiding assumptions about the unknown noise-free signal. Furthermore, the spatial dependence of the quantum noise, the electronic noise and the noise spatial correlation were taken into account in the simulation, which enabled the generation of accurate clinical image samples for each patient, as if they had been acquired with lower radiation doses. At the end, the assessment of the trained model was performed on real LD mammography acquisitions using a physical anthropomorphic breast phantom.

The main contributions of this paper are summarized as follows.

- A comprehensive evaluation of common loss functions specifically to the field of medical imaging with an emphasis on digital mammography.
- A new training strategy using 400 LD clinical data from retrospective mammography examinations was proposed. This enables the model to have contact with a diverse dataset, with different tissue structures, densities, signal and noise levels, etc.
- Validation of each loss function with an objective metric that can evaluate blurring and noise individually.

The remainder of the paper is organized as follows. Section 2.3 introduces the theoretical background for the image degradation model and for the restoration model including both the MB and data-based, and the loss functions used in this study. Section 2.4 presents the datasets used in this work, all the implementation details and the metrics used for the evaluation. In Section 2.5, the quantitative results and also some regions of interest (ROIs) are shown for all the loss functions, followed by a concluding summary in Section 2.6.

2.3 Theoretical background

The image restoration methods can be approached on several fronts. A common practice, which has been extensively presented in the literature, is to model the x-ray acquisition system and create mathematical formulations to perform the desired restoration. Although this was a common procedure in the past few years, new deep learning techniques have shown a great capacity of learning from the data, through supervised learning. This section presents the background information behind the image degradation model for mammography systems, the basics of the MB approach to restore the LD images and the used model.

2.3.1 Degradation model

Considering $\boldsymbol{X} \in \mathbb{R}^{w \times h}$ an observed X-ray mammography image of size $w \times h$ at standard FD, following (BORGES *et al.*, 2018), we can model an acquisition as follows:

$$\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{\eta}, \quad \text{s.t.} \quad [\boldsymbol{\eta}]_{ij} \sim \mathcal{N}\left(0, \lambda \boldsymbol{Y}_{ij} + \sigma_e^2\right), \tag{2.1}$$

where $\boldsymbol{Y}$ is the noise-free image, and $\boldsymbol{\eta}$ is the corresponding noise at standard FD. Note that $[\boldsymbol{\eta}]_{ij}$ represents the element at the i-th row and j-th column of the noise matrix $\boldsymbol{\eta}$, and follows a Gaussian distribution with zero mean and variance equal to $\lambda \boldsymbol{Y}_{ij} + \sigma_e^2$; here, λ is the quantum noise gain and σ_e^2 is the variance of the electronic noise. Although in x-ray images the noise is often modeled through a Poisson-Gaussian distribution, the energy ranges at which digital mammography operates allow the assumption of a signal-dependent Gaussian distribution, as done in Eq. (2.1), thanks to the Central Limit Theorem (AZZARI; BORGES; FOI, 2018). Now let us consider another raw mammogram $\boldsymbol{X}_\gamma$, acquired using the same radiographic factors as $\boldsymbol{X}$ except for a reduction in current-time product (mAs), resulting in lower dose. The mammogram $\boldsymbol{X}_\gamma$ can be described as a function of the noise-free signal in Eq. (2.1) as follows:

$$\boldsymbol{X}_\gamma = \gamma \boldsymbol{Y} + \boldsymbol{\eta}_\gamma, \quad \text{s.t.} \quad [\boldsymbol{\eta}_\gamma]_{ij} \sim \mathcal{N}\left(0, \gamma \lambda \boldsymbol{Y}_{ij} + \sigma_e^2\right), \tag{2.2}$$

where $0 < \gamma < 1$ is the mAs scaling factor, and thus the dose reduction factor.

The goal of this work is to investigate appropriate loss-functions capable of training a CNN to achieve Eq. (2.1) starting from Eq. (2.2), while keeping the noise-free signal $\boldsymbol{Y}$ as preserved as possible, with minimal blur and smear caused by the restoration process, *i.e.*:

$$\widehat{\boldsymbol{X}} = \Psi(\boldsymbol{X}_\gamma), \tag{2.3}$$

where $\widehat{\boldsymbol{X}}$ is the restored image and $\Psi(\cdot)$ is the non-linear restoration operator.

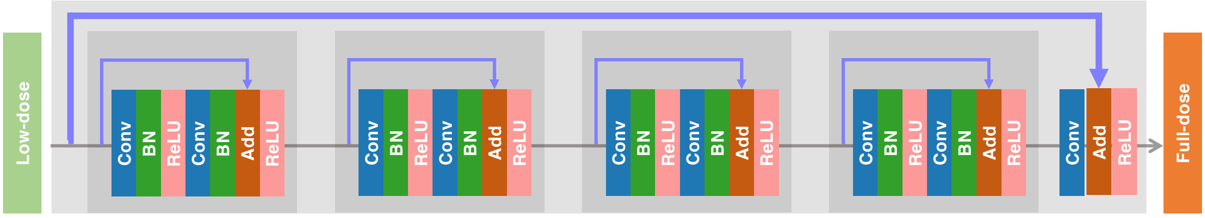


Figure 1 – Architecture of the network used in this study.

2.3.2 Restoration model

From Eqs. (2.1) and (2.2), the restoration task may be approached as a mathematical operator, which we will refer to as MB approaches. Alternatively, taking advantage of the great capacity of DNN, it is also possible to train a model to perform the whole restoration process avoiding the estimation of noise parameters. In this section we discuss both the model- and data-based approaches.

2.3.2.1 Model-based

In (BORGES *et al.*, 2017; BORGES *et al.*, 2018), our group proposed a MB pipeline to restore LD mammography images, leveraging a variance stabilization technique, namely the generalized Anscombe transformation (GAT) (STARCK; MURTAGH; BIJAOUI, 1998). With this technique, it is possible to use any denoising technique designed to treat signal-independent Gaussian-distributed data. The pipeline involves modeling the equipment's noise parameters, such as quantum noise gain, electronic noise variance, and the detector offset. These components are used in the GAT to bring the image to the domain where the noise model is signal-independent and approximately Gaussian. In the GAT domain, the block-matching and 3D filtering (BM3D) (DABOV *et al.*, 2007) is used to suppress noise and the exact unbiased inverse of the generalized Anscombe transform is applied (MAKITALO; FOI, 2012). Finally, the denoised image is blended with the LD image through a weighted average. We refer to Borges *et al.* (2018) for more details about the MB restoration process.

2.3.2.2 Data-based

This subsection presents the deep learning-based denoising model for restoration of LD digital mammography. Despite the high dimensionality of the Euclidean space in medical images, it is known that these images lie in low-dimensional manifolds (WU *et al.*, 2017; YANG *et al.*, 2018a). As DNNs, given sufficient trainable parameters, can approximate any non-linear transformation functions, one can model the non-linear transformation Ψ, from Eq. (2.3), using deep neural-networks with the appropriate architecture and loss-function.

Inspired by the success of residual skip connection in various tasks such as image classification (HE *et al.*, 2016), image denoising (WOLTERINK *et al.*, 2017; CHEN *et al.*, 2017a), reinforcement learning (SILVER *et al.*, 2017), we used the ResNet to better model the noise distribution of the LD digital mammography.

The network has four residual blocks, each of them has a skip connection and serves as basic units. Each residual block contains four layers: two convolutional layers and two batch-normalization layers (IOFFE; SZEGEDY, 2015). The batch normalization layer has been proved to accelerate deep network training by reducing internal covariate shift. Rectified linear unit (ReLU) activation function is used after batch normalization layer or addition operation.

Specifically, all convolutional layers have 64 convolutional filters of size 3 × 3 with a stride of 1 and a zero-padding of 1 except for the final convolutional layer that has only one convolutional filter. Throughout the network, the feature-maps have the same size as the input image.

2.3.3 Loss functions

The loss function plays an important role in training a restoration model and can determine the visual aspect of the generated images. In Zhao *et al.* (2016), the authors investigated the effects of several commonly-used losses for image restoration with neural networks. The difference between this study and theirs lies in two aspects. First, we study the effects of different losses for image restoration of LD digital mammography, while Zhao *et al.* (2016) focused on natural images. In digital mammography, as opposed to natural images, specific high-frequency and low-contrast features of the exams, such as MC and small masses, are vital for the successful clinical use of the data. Thus, in this scenario, it is preferable to retain some residual noise and keep those subtle features intact instead of aggressively filtering the data at the risk of masking lesions. Second, in addition to those metrics studied in Zhao *et al.* (2016), we also studied adversarial loss, the PL (JOHNSON; ALAHI; FEI-FEI, 2016), which is based on a pretrained VGG model (SIMONYAN; ZISSERMAN, 2014) and the FID (MATHIASEN; HVILSHØJ, 2020; HEUSEL *et al.*, 2017), which is based on the Inception model (SZEGEDY *et al.*, 2015).

Different losses compute the similarity between the generated and GT images in different ways. We denote the generated and GT images as $\widehat{\boldsymbol{X}} \in \mathbb{R}^{w \times h}$ and $\boldsymbol{X} \in \mathbb{R}^{w \times h}$, respectively.

2.3.3.1 Mean squared error

MSE is the most widely used metric to measure the pixel-wise difference between generated and GT images. It can be formally defined as:

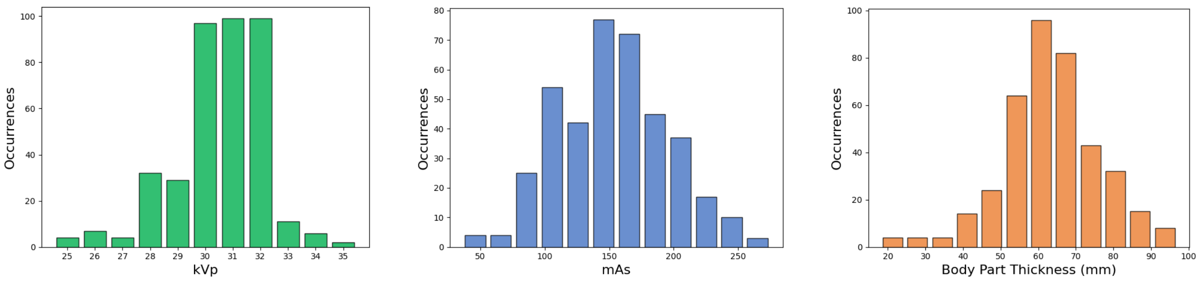
$$\mathcal{L}_{\text{MSE}} = \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \left([\widehat{\boldsymbol{X}}]_{ij} - [\boldsymbol{X}]_{ij} \right)^2 , \tag{2.4}$$

where $\boldsymbol{X}_{ij}$ indicates the element at i-th row and j-th column in $\boldsymbol{X}$.

2.3.3.2 Mean absolute error or ℓ_1

Slightly different from MSE, MAE computes the ℓ_1 loss between the generated and GT image. As a result, MAE does not over-penalize larger errors and can overcome the smoothness caused by MSE. It can be defined as:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \left| [\widehat{\boldsymbol{X}}]_{ij} - [\boldsymbol{X}]_{ij} \right| . \tag{2.5}$$



(a) (b) (c)

Figure 2 – Histograms showing the variability of (a) kVp, (b) mAs and (c) breast thickness in the clinical images from the training dataset.

2.3.3.3 Structural similarity index

SSIM is a widely-used image quality evaluation metric (WANG *et al.*, 2004). SSIM can measure the visual similarity between two images in terms of their structures and textures. The SSIM index is computed based on various windows of an image. The measure between the window $\widehat{\boldsymbol{x}}$ over $\widehat{\boldsymbol{X}}$ and the window $\boldsymbol{x}$ over $\boldsymbol{X}$, based on a common window size $k \times k$, can be defined as:

$$\text{SSIM}(\widehat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{2\mu_{\widehat{\boldsymbol{x}}}\mu_{\boldsymbol{x}} + c_1}{\mu_{\widehat{\boldsymbol{x}}}^2 + \mu_{\boldsymbol{x}}^2 + c_1} \frac{2\sigma_{\widehat{\boldsymbol{x}}\boldsymbol{x}} + c_2}{\sigma_{\widehat{\boldsymbol{x}}}^2 + \sigma_{\boldsymbol{x}}^2 + c_2}, \tag{2.6}$$

where $\mu_{\widehat{\boldsymbol{x}}}$ and $\mu_{\boldsymbol{x}}$ are the averages of $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}$ respectively, $\sigma_{\widehat{\boldsymbol{x}}}$ and $\sigma_{\boldsymbol{x}}$ are the variances of $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}$ respectively, and $\sigma_{\widehat{\boldsymbol{x}}\boldsymbol{x}}$ is the covariance of $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}$. Also, $c_1 = 1 \times 10^{-4}$ and $c_2 = 9 \times 10^{-4}$ are two constants, which are used to stabilize the division with a weak denominator. The window size k is 11, as suggested in Wang *et al.* (2004). The MSSIM between two images $\widehat{\boldsymbol{X}}$ and $\boldsymbol{X}$, $\text{MSSIM}(\widehat{\boldsymbol{X}}, \boldsymbol{X})$, refers to the average of the SSIM index over all windows. The SSIM loss is defined as:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{MSSIM}\left(\widehat{\boldsymbol{X}}, \boldsymbol{X}\right) . \tag{2.7}$$

2.3.3.4 Perceptual loss

PL attempts to compare the similarity between two images in a high-level feature space (JOHNSON; ALAHI; FEI-FEI, 2016). A pretrained VGG model is widely used to extract features from an image to form such a high-level feature space, which is expected to mimic the human visual system. PL is similar to MSE, but in a feature space instead of pixel space. It can be defined as:

$$\mathcal{L}_{\text{PL}} = \frac{1}{w'h'c'} \sum_{i=1}^{w'} \sum_{j=1}^{h'} \sum_{k=1}^{c'} \left([\boldsymbol{\Phi}(\widehat{\boldsymbol{X}})]_{ijk} - [\boldsymbol{\Phi}(\boldsymbol{X})]_{ijk} \right)^2, \tag{2.8}$$

where $\boldsymbol{\Phi}$ represents the feature extractor, whose output is a tensor of size $w' \times h' \times c'$. The PL can be computed on early or later layers of the VGG network. Each layer is commonly denominated as a block of convolutions and activation functions, *e.g.* ReLU, before the max-pooling. For example, PL1 contains the first two convolution layers and their respective activation function.

2.3.3.5 Adversarial loss

Adversarial loss was first presented in (GOODFELLOW *et al.*, 2014) to train a model to generate realistic synthetic images. The GAN framework contains two networks: a generator (G) to create the images and a discriminator (D) to evaluate the created image. In this work, we used the Wasserstein GANs with gradient penalty (WGAN-GP) (GULRAJANI *et al.*, 2017). The generator loss is given by:

$$\mathcal{L}_{\text{GAN}} = \mathcal{L}_{\text{MAE}} + \lambda_{adv} * \mathcal{L}_{\text{adv(G)}}, \tag{2.9}$$

and

$$\mathcal{L}_{\text{adv(G)}} = -\,\mathbb{E}\{D(\widehat{\boldsymbol{X}})\}, \tag{2.10}$$

where $\mathcal{L}_{\text{MAE}}$ is the pixel-wise previously stated, λ_{adv} is weighting factor, and $\mathcal{L}_{\text{adv(G)}}$ the adversarial loss for the generator. We followed the the discriminator loss is given by the original paper (GULRAJANI *et al.*, 2017).

2.3.3.6 Fréchet inception distance

FID was first proposed to measure the image quality of images generated by GANs (HEUSEL *et al.*, 2017). The metric is calculated as:

$$\mathcal{L}_{\text{FID}} = ||\mu_{\widehat{\boldsymbol{x}}} - \mu_{\boldsymbol{x}}||^2 + \text{Tr}(\Sigma_{\widehat{\boldsymbol{x}}} + \Sigma_{\boldsymbol{x}} - 2 * \sqrt{\Sigma_{\widehat{\boldsymbol{x}}} * \Sigma_{\boldsymbol{x}}}), \tag{2.11}$$

where $\text{Tr}(\cdot)$ is the trace of a matrix, μ is the mean vector, and Σ the covariance matrix. Both μ and Σ are calculated over feature map vectors acquired after a batch of images are forwarded through a Inception-v3 model.

2.4 Materials & methods

For reliable training and validation, we created two distinct sets of images: the first with clinical cases and the second with anthropomorphic breast phantom images. We used a hybrid dataset of clinical images with simulated LD/FD pairs and later restricted testing the trained model in the phantom data, with LD acquisitions acquired directly in the mammography equipment. Doing so, we avoid the so-called "inverse crime", which is to test and train the model with the same fabricated synthetic data (ZHAO *et al.*, 2016). In this section, we specify how both datasets used in this study were constructed. Also, in this section we present the DNN implementation details and the evaluation metrics used in this study.

2.4.1 Training dataset

The training dataset consists of 400 clinical mammography acquired at the Barretos Cancer Hospital (Brazil). These data were obtained retrospectively from breast cancer screening examinations, after approval by the institutional review board. This dataset is relative to 100 patients with their respective images from the craniocaudal (CC) and mediolateral oblique (MLO) views of the left and right breast. All images were acquired using a Hologic Selenia Dimensions Mammography System (Hologic, Bedford, MA) and all the images were saved as raw data, *i.e.*, DICOM "*for processing*". Fig. 2 shows the occurrence of different values of kVp, mAs, and breast thickness in the training dataset. All clinical images were fully anonymized to preserve patients' medical records.

As previously discussed, exposing patients to x-ray radiation several times to build an image dataset with LD acquisitions can be dangerous and impractical in the clinical routine. In order to acquire clinical images at lower doses, we used a previous work to simulate dose reduction in these data (BORGES *et al.*, 2016; BORGES *et al.*, 2017). The method injects quantum and electronic noise in the VST domain and also accounts for the detector crosstalk. We refer to both these works for further details about this technique. We applied this technique to all clinical images to simulate acquisitions of 75% and 50% of the standard FD, in which the image was originally obtained ($\gamma = 0.75$ and $\gamma = 0.50$, respectively). After the simulation, we reached a total of 1,200 images among full and reduced doses.

2.4.2 Testing dataset

To validate the restoration methods that were trained on the clinical dataset, we acquired images of a physical anthropomorphic breast phantom at the Hospital of the University of Pennsylvania. We also used a Hologic Selenia Dimensions Mammography System, as for the training dataset. The phantom has six slabs, with total thickness of 51 mm. It consists of a material that mimics the breast tissue and was prototyped by CIRS, Inc. (Reston, VA), under the license from the University of Pennsylvania (CARTON *et al.*, 2011). Small pieces of calcium oxalate (99%, Alfa Aesar, Ward Hill, MA) were placed between the six slabs to simulate MCs, as illustrated in Fig. 3.

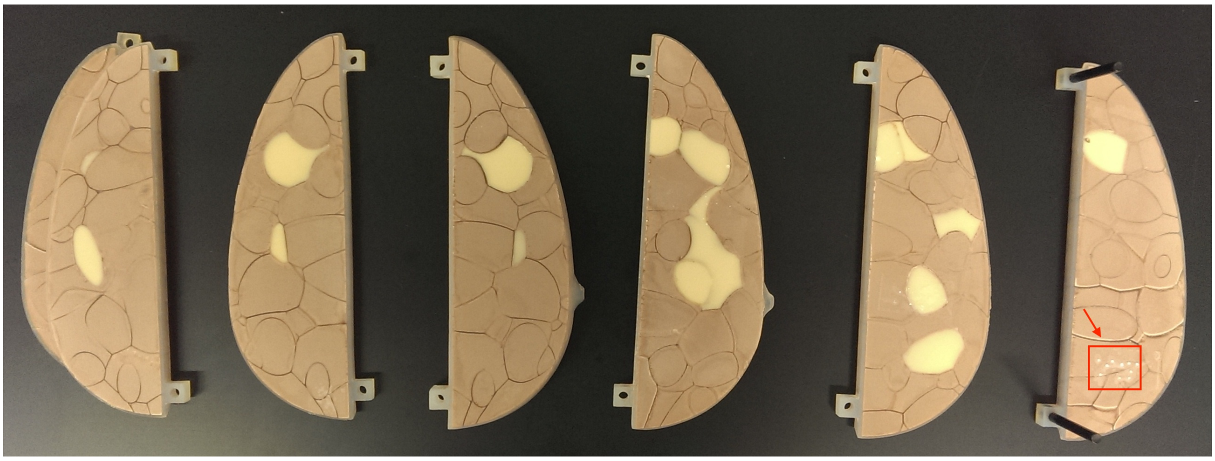


Figure 3 – Image of the physical anthropomorphic breast phantom used in this study. The red arrow points to the pieces of calcium oxalate that were placed between the slabs to simulate a cluster of microcalcifications.

We acquired a total of 25 images of the anthropomorphic breast phantom. Firstly, for the FD, the automatic exposure control (AEC) was used to select the standard radiographic factors, which yielded a combination of 29 kVp and 160 mAs. Without physically moving the phantom, the system was set to manual mode and 15 exposures at the standard radiographic factors were performed to generate a set of FD images. For reduction factors of 75% and 50%, 10 images were acquired by reducing the current-time product from 160 mAs to 120 mAs and 80 mAs respectively. It is important to note that all these images were saved as raw data.

2.4.3 Figures of merit

To take advantage of being capable of exposing the physical anthropomorphic breast phantom several times, we generated a pseudo-GT and compared all the restored images with this GT through an assessment of the signal-to-noise ratio (SNR) and the mean normalized squared error (MNSE) decomposed into residual noise ($\mathcal{R}_{\mathcal{N}}$) and bias squared ($\mathcal{B}^2$). This metric was previously presented in Borges *et al.* (2018) for the assessment of digital breast tomosynthesis images.

We first measured the signal-to-noise ratio (SNR) in all phantom images and its restorations as the ratio of the mean pixel value and its standard deviation along the five realizations. The metric was calculated only inside the breast region and an average filter of size 15×15 was used to smooth both the mean signal value and the standard deviation after their calculation.

Considering $\mathcal{X} \subset \mathbb{R}^{w \times h}$ a subspace of FD mammography acquisitions, suppose we have a set of N realizations $\boldsymbol{X}^* \in \mathcal{X}$ for the the GT estimation, and a set of P realizations $\boldsymbol{X}^{'} \in \mathcal{X}$ for MNSE assessment. Here we refer to the GT as the expectation of

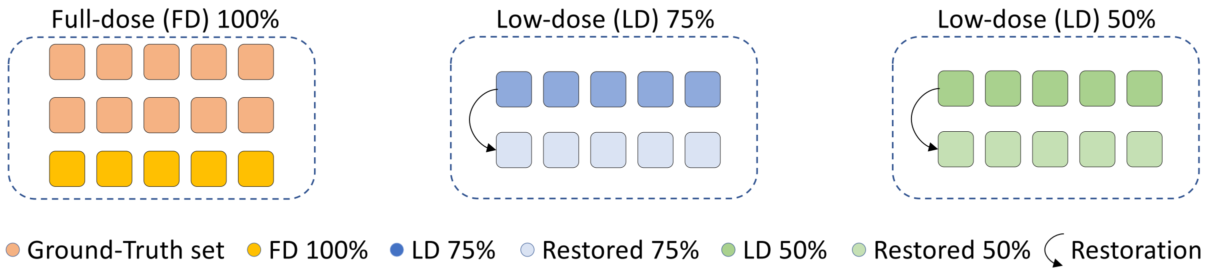


Figure 4 – Illustration of how the anthropomorphic breast phantom dataset was subdivided for the assessment of the MNSE and its decomposition. The metrics were evaluated on all different colored sub-groups of FD, LD and restored images.

the acquisitions on this subspace $\widehat{\boldsymbol{X}}_{GT} = \mathbb{E}\{\boldsymbol{X}^*\}$. Also, after breast phantom segmentation, we denote (i, j) as a pair of 2D indices running inside a set $\mathcal{I} \in \mathbb{X}^+$ of size $\boldsymbol{I} < w \times h$. The set $\mathcal{I}$ represents the collection of pixel coordinates after the segmentation. The MNSE can be calculated as:

$$\text{MNSE} = \frac{1}{P}\sum_{p=1}^{P} \underbrace{\frac{1}{\boldsymbol{I}} \sum_{(i,j)\in\mathcal{I}} \frac{\left([\boldsymbol{X}'_p]_{ij} - [\widehat{\boldsymbol{X}}_{GT}]_{ij}\right)^2}{[\widehat{\boldsymbol{X}}_{GT}]_{ij}}}_{\text{NQE}} - \underbrace{\frac{1}{\boldsymbol{I}N} \sum_{(i,j)\in\mathcal{I}} \frac{\mathbb{V}\left([\boldsymbol{X}^*]_{ij}\right)}{[\widehat{\boldsymbol{X}}_{GT}]_{ij}}}_{\phi_1}, \tag{2.12}$$

where ϕ_1 is accounting for the error associated with the limited number of images used for the GT estimation and $\mathbb{V}\left(\boldsymbol{X}^*\right)$ is a point-wise variance among this set of images. It is possible to decompose the MNSE into $\mathcal{R}_\mathcal{N}$ and $\mathcal{B}^2$ portions, such that:

$$\text{MNSE} = \underbrace{\left\{ \frac{1}{\boldsymbol{I}} \sum_{(i,j)\in\mathcal{I}} \frac{\left(\mathbb{E}\{[\boldsymbol{X}']_{ij}\} - [\widehat{\boldsymbol{X}}_{GT}]_{ij}\right)^2}{[\widehat{\boldsymbol{X}}_{GT}]_{ij}} \right\} - \phi_1 - \frac{\mathcal{R}_\mathcal{N}}{P}}_{\mathcal{B}^2} + \underbrace{\frac{1}{\boldsymbol{I}} \sum_{(i,j)\in\mathcal{I}} \frac{\mathbb{V}\left([\boldsymbol{X}']_{ij}\right)}{[\widehat{\boldsymbol{X}}_{GT}]_{ij}}}_{\mathcal{R}_\mathcal{N}}. \tag{2.13}$$

For the practical evaluation, we first manually segmented the phantom image to avoid calculating the metric outside the breast tissue. From the 15 images at the FD, 10 were used to generate the GT ($N = 10$) and 5 to calculate the metric ($P = 5$). As previously mentioned, 5 images of each reduced dose ($P_\gamma = 5$) were used to calculate the metric as well. The scheme shown in Fig. 4 illustrates how we separated all the phantom images. The MNSE and its decomposition were evaluated on all different colored sub-groups of LD and restored images.

To avoid error on the $\mathcal{B}^2$ due to differences in the mean value from different acquisitions, we used a technique of fitting a first-order polynomial to correct the image mean value. First, the GT was calculated following the previous equations, and then all the images used to generate the GT itself were adjusted based on the calculated GT, as done in Borges *et al.* (2018). After this correction, the GT was calculated again and all the non-GT FD acquisitions were also adjusted through the fitting technique. All the restored images and the LD acquisitions had their mean value adjusted through this method. This overall process guarantees that the $\mathcal{B}^2$ error measurement is due to the smearing/blurring imposed by the restoration process and not small changes in the mean value of the images.

We also compared the time spent to run the restoration process on the model based approach and in the deep network.

2.4.4 Implementation details

Overall, we trained 22 models in this study: 11 dedicated to restore 75%-dose images and 11 dedicated to 50%-dose. Each of these 11 models refer to the mentioned loss functions from Section 2.3.3. We also performed 3-fold cross-validation for each loss, leading to a total of 66 models. Although there is a range of different loss functions for image restoration, as shown in (DING *et al.*, 2021), we used in this work examples of error visibility methods, *e.g.*, the MAE; structural similarity method, *e.g.*, the SSIM; and DNN method, *e.g.*, the PL, FID, and adversarial loss. These losses are commonly used by previous work in the field of medical imaging restoration (CHEN *et al.*, 2017a; CHEN *et al.*, 2017b; YANG *et al.*, 2018a; KANG *et al.*, 2018; SHAN *et al.*, 2018; YANG *et al.*, 2018b; SHAN *et al.*, 2019). For the PL, following Johnson, Alahi and Fei-Fei (2016), we used the VGG-16 network (SIMONYAN; ZISSERMAN, 2014), which was pretrained on the ImageNet dataset and publicly available at the PyTorch official website. Since different layers of the VGG-16 networks can form different feature spaces, we used the feature-maps right before the first four max-pooling layers to form four different feature spaces, and then obtain four corresponding PLs from shallow to deep. We denote these four PLs as $\mathcal{L}_{\text{PL1}}$, $\mathcal{L}_{\text{PL2}}$, $\mathcal{L}_{\text{PL3}}$, and $\mathcal{L}_{\text{PL4}}$. We found that too many max-pooling layers involved in PL largely affect pixel-wise comparison, therefore, we removed all max-pooling layers in PL4. For the FID loss, we used a pre-trained Inception v3 model and followed the work of Mathiasen and Hvilshøj (MATHIASEN; HVILSHØJ, 2020) for implementing the algorithm[1]. In the adversarial loss, we use three candidate values for λ_{adv} to illustrate the different impacts of this parameter. We used the ResNet as architecture for the generator and the same architecture described in the original paper for the discriminator (also called critic).

For all neural networks with different losses, the initial learning rate λ was set as 1.0×10^{-4} and was reduced to half every 10 epochs. The trainable parameters in the network were optimized using the Adam optimization (KINGMA; BA, 2014), whose coefficients used for computing running averages of gradients and its square were set as 0.5 and 0.999, respectively. The network was implemented with PyTorch DL library (PASZKE *et al.*, 2017) and trained within 60 epochs using a NVIDIA GeForce GTX 1080 Ti GPU. The batch size was set as 256 during the training; however, it was reduced accordingly for the training with PL since VGG network also needs to be in the GPU. To make the training for non-pixel-wise losses easier, we used a pre-trained MAE network as the starting

[1] Available at www.github.com/AlexanderMath/FastFID

point.

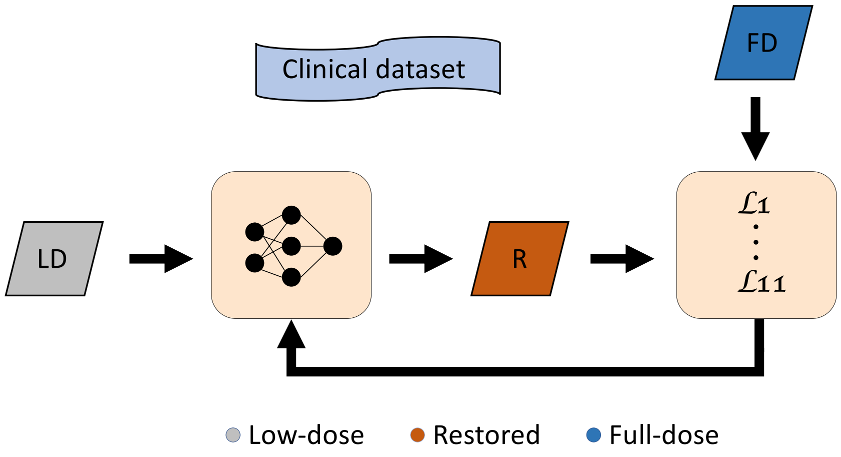
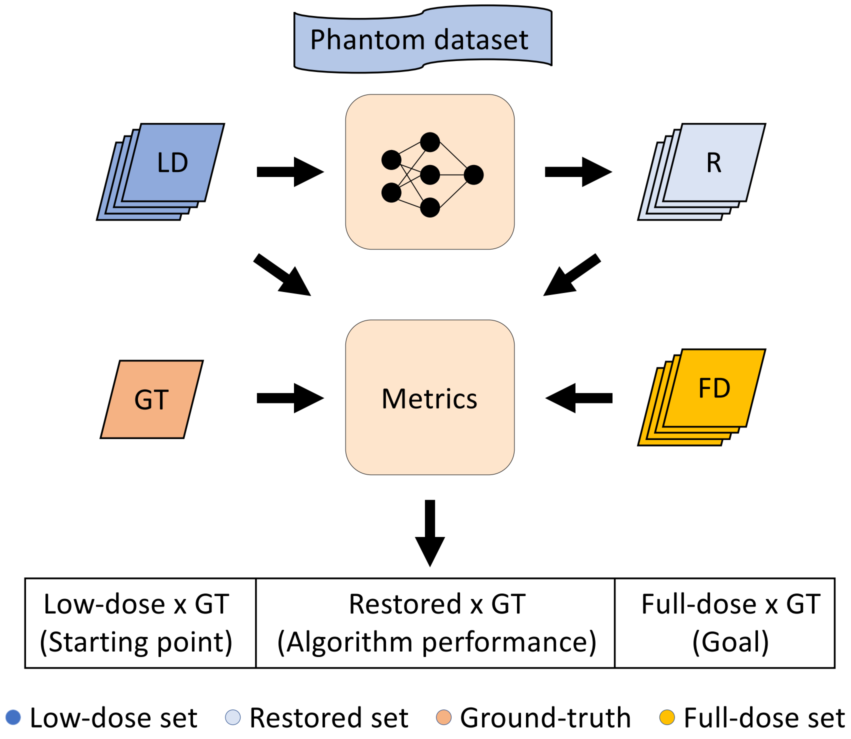


Figure 5 – Schematic explaining the training process of the models.



Low-dose x GT (Starting point)	Restored x GT (Algorithm performance)	Full-dose x GT (Goal)

Figure 6 – Schematic explaining the test procedure for the models.

To train the restoration model, a total of 256,000 patches of size 64×64 were randomly selected from the breast regions of the 400 clinical images available in this study. These patches include pairs of LD and FD images, *i.e.*, we trained one neural network for each reduction factor.

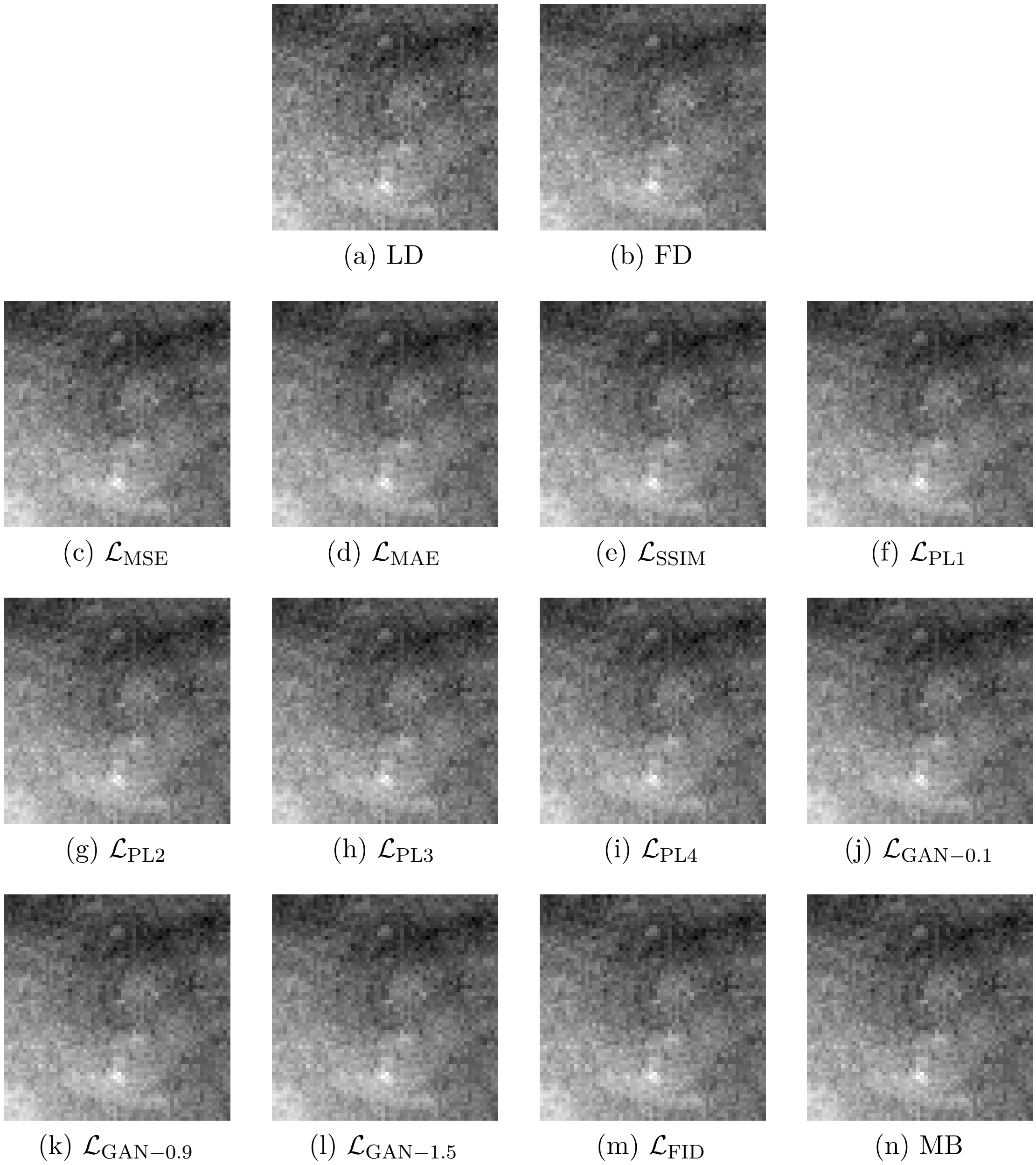
After the training process, which is illustrated in Fig. 5, the model was quantitatively tested on the anthropomorphic breast phantom dataset to validate the effects of different loss functions. The testing stage, illustrated in Fig. 6, involved measuring the MNSE between the GT against the LD, FD and restored images. The value from the LD×GT gives us the starting point of $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$. At this stage, $\mathcal{R}_{\mathcal{N}}$ is usually high and $\mathcal{B}^2$ low. The FD×GT guide us to the goal of the restoration regarding the lowest $\mathcal{B}^2$ and the desired $\mathcal{R}_{\mathcal{N}}$ level. When evaluating the Restored×GT, it is possible to see how the restoration performed, *i.e.*, how much blurring was imposed, measuring through the $\mathcal{B}^2$, and how close the $\mathcal{R}_{\mathcal{N}}$ is to the FD value. Implementation details of how this metric was calculated on each of these groups are presented in Section 2.4.3.

2.5 Results & discussions

In this section, we present and discuss the quantitative assessment performed on the phantom images and also the ROIs for both clinical and phantom images. We compared the LD and FD acquisitions to the images restored by the model using the following loss functions: MSE, MAE, SSIM, PL1, PL2, PL3, PL4, GAN, and FID. Also, for comparison with an analytical method previously published in the literature, we used the model-based (MB) approach proposed in (BORGES *et al.*, 2018) as a benchmark for all different loss functions.

2.5.1 Visual analysis

Figs. 7 and 8 display a magnified ROI of a patient image acquired at the standard radiation dose and at the simulated dose reduction factors of 75% and 50%, respectively. For all visual analysis results, we show the results of each loss function that achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD within all training realizations. As expected, the LD mammography presents more perceived noise as compared with the FD mammography. The restored results of LD mammography are shown in (c)-(j). We chose an ROI with a MC cluster, which is an important feature for cancer diagnosis. As expected, for the reduction factor of 50% (Fig. 8), we can see that the loss functions MSE and MAE yield an overall smoothing in the image. For the SSIM, there is a subtle difference in the noise level when compared with MAE and MSE. Again, when we go from PL1 to PL4, it is visually noticeable that the blurring effect decreases and the fine details are preserved. Moreover, we can recognize that PL3 has slightly more noise compared with PL4, which can be confirmed in the objective metrics, specifically in the $\mathcal{R}_{\mathcal{N}}$ in the MNSE decomposition. We can see that FID also achieves good results with more overall noise compared with PL3 and PL4. This is also observed in the objective metrics as expected since this loss uses a pre-trained network to extract image features. The adversarial losses could also restore the image with less noise compared with the losses that use DNN as image quality metric. It is important to note the visual similarity between the DNN with FID, PL3 and PL4 loss functions and the MB method. All the visual analysis discussed agree with the quantitative results presented in the next section. The same discussion can be done for the 75% case, in Fig. 7; however, the differences are more subtle when compared with the 50%.



(a) LD (b) FD

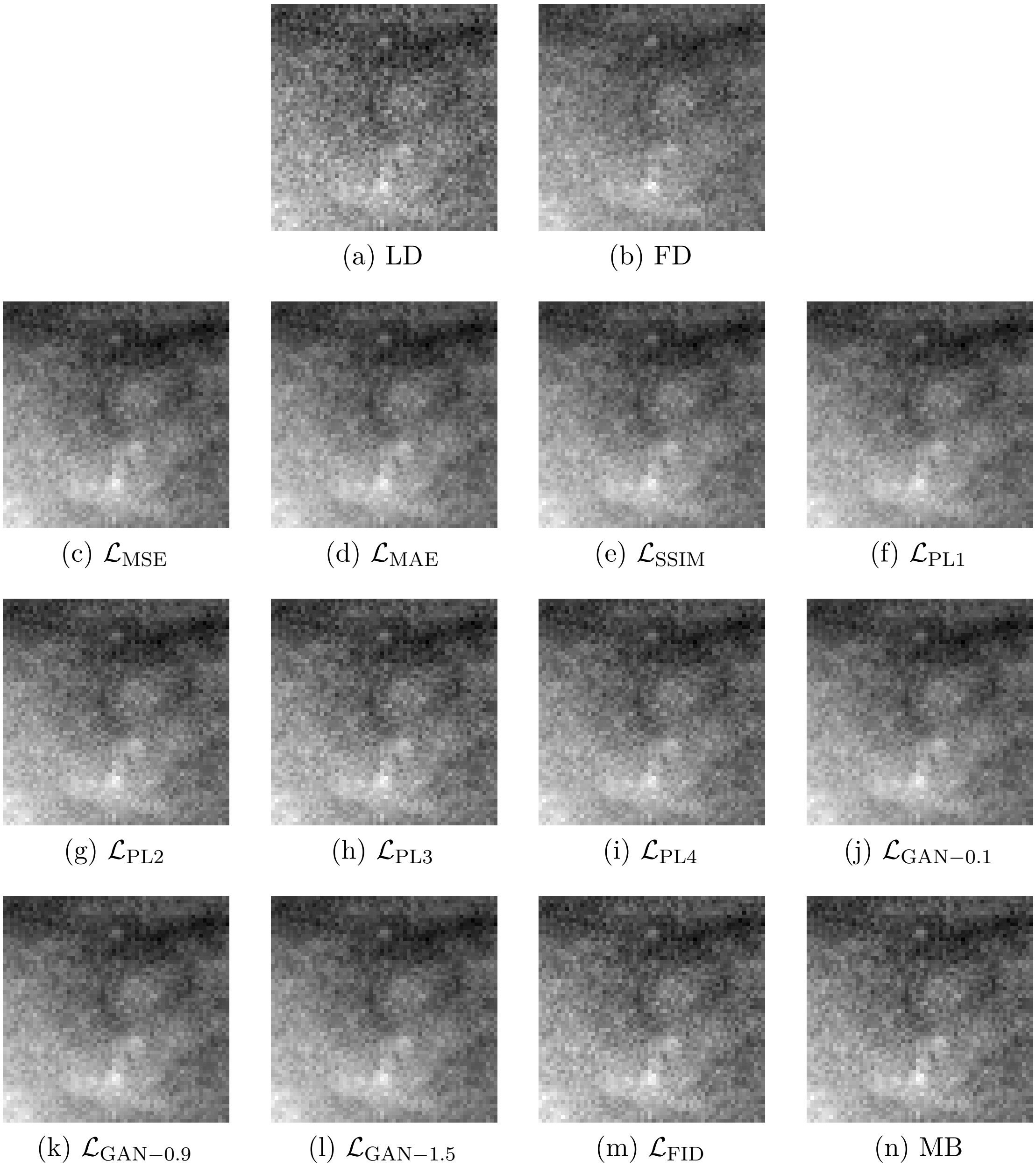
(c) $\mathcal{L}_{\text{MSE}}$ (d) $\mathcal{L}_{\text{MAE}}$ (e) $\mathcal{L}_{\text{SSIM}}$ (f) $\mathcal{L}_{\text{PL1}}$

(g) $\mathcal{L}_{\text{PL2}}$ (h) $\mathcal{L}_{\text{PL3}}$ (i) $\mathcal{L}_{\text{PL4}}$ (j) $\mathcal{L}_{\text{GAN}-0.1}$

(k) $\mathcal{L}_{\text{GAN}-0.9}$ (l) $\mathcal{L}_{\text{GAN}-1.5}$ (m) $\mathcal{L}_{\text{FID}}$ (n) MB

Figure 7 – Illustration of a magnified ROI of a clinical image focusing a small cluster of MC. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv} and (m) FID. (n) model-based restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range.

Figs. 9 and 10 display a magnified ROI from the mammography image of the anthropomorphic breast phantom at dose reduction factors of 75% and 50%, respectively. In this case, it is important to note that the radiation dose reduction was performed on the equipment itself, changing the mAs with each acquisition. The same discussion previously mentioned for the clinical images can be used here as the images have the same visual



(a) LD

(b) FD

(c) $\mathcal{L}_{\text{MSE}}$

(d) $\mathcal{L}_{\text{MAE}}$

(e) $\mathcal{L}_{\text{SSIM}}$

(f) $\mathcal{L}_{\text{PL1}}$

(g) $\mathcal{L}_{\text{PL2}}$

(h) $\mathcal{L}_{\text{PL3}}$

(i) $\mathcal{L}_{\text{PL4}}$

(j) $\mathcal{L}_{\text{GAN}-0.1}$

(k) $\mathcal{L}_{\text{GAN}-0.9}$

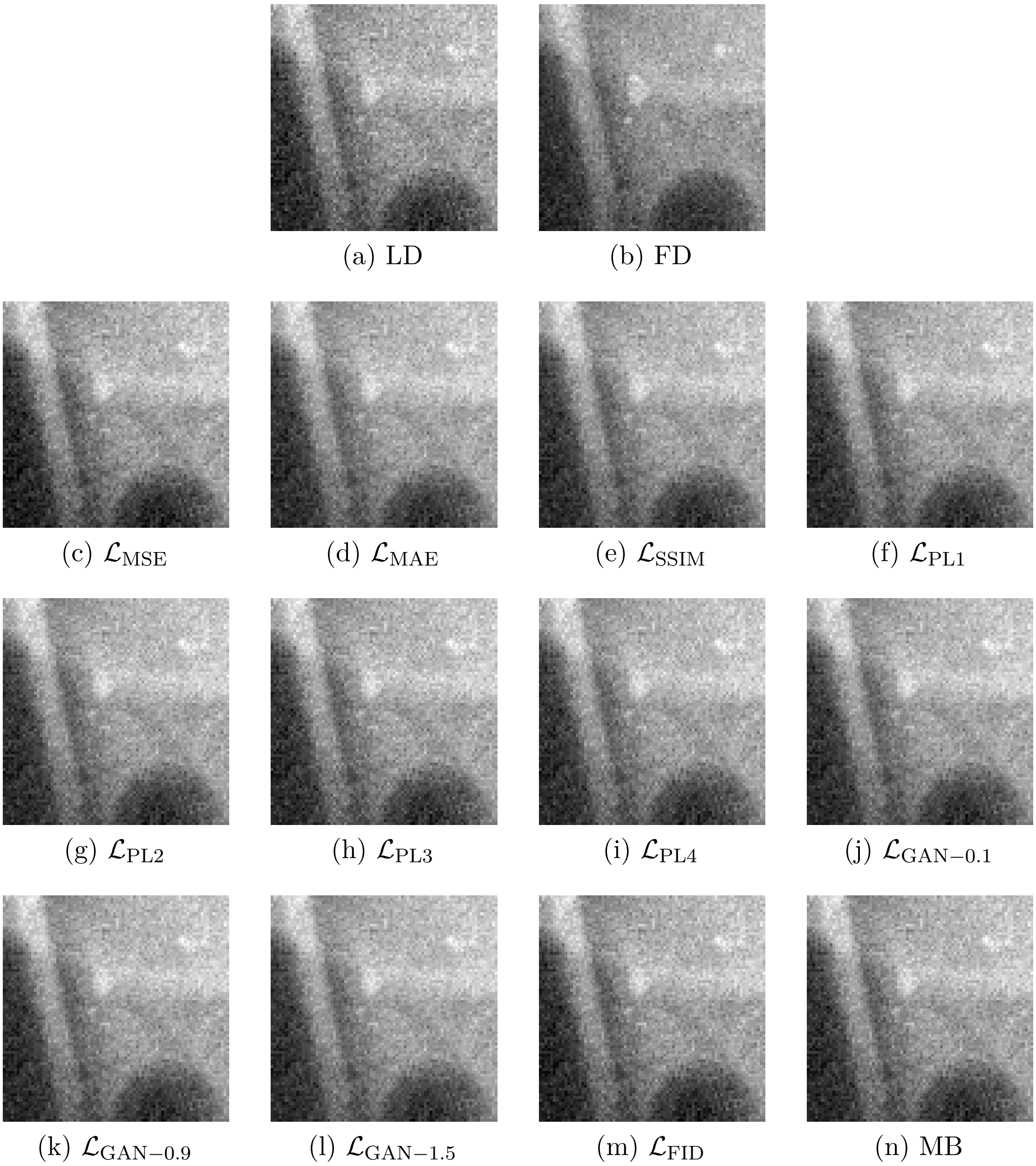
(l) $\mathcal{L}_{\text{GAN}-1.5}$

(m) $\mathcal{L}_{\text{FID}}$

(n) MB

Figure 8 – Illustration of a magnified ROI of a clinical image focusing a small cluster of MC. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv} and (m) FID. (n) model-based restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range.

properties. From the results of the anthropomorphic phantom we can infer that the neural network is generalizing well even for cases with slightly different noise properties, *i.e.*, even with a careful simulation as used in this work some discrepancies are expected between the simulated LD image and the actual LD image; and the trained model was able to overcome such small discrepancies. Also, it reinforces the fact that the model might be



(a) LD

(b) FD

(c) $\mathcal{L}_{\text{MSE}}$

(d) $\mathcal{L}_{\text{MAE}}$

(e) $\mathcal{L}_{\text{SSIM}}$

(f) $\mathcal{L}_{\text{PL1}}$

(g) $\mathcal{L}_{\text{PL2}}$

(h) $\mathcal{L}_{\text{PL3}}$

(i) $\mathcal{L}_{\text{PL4}}$

(j) $\mathcal{L}_{\text{GAN}-0.1}$

(k) $\mathcal{L}_{\text{GAN}-0.9}$

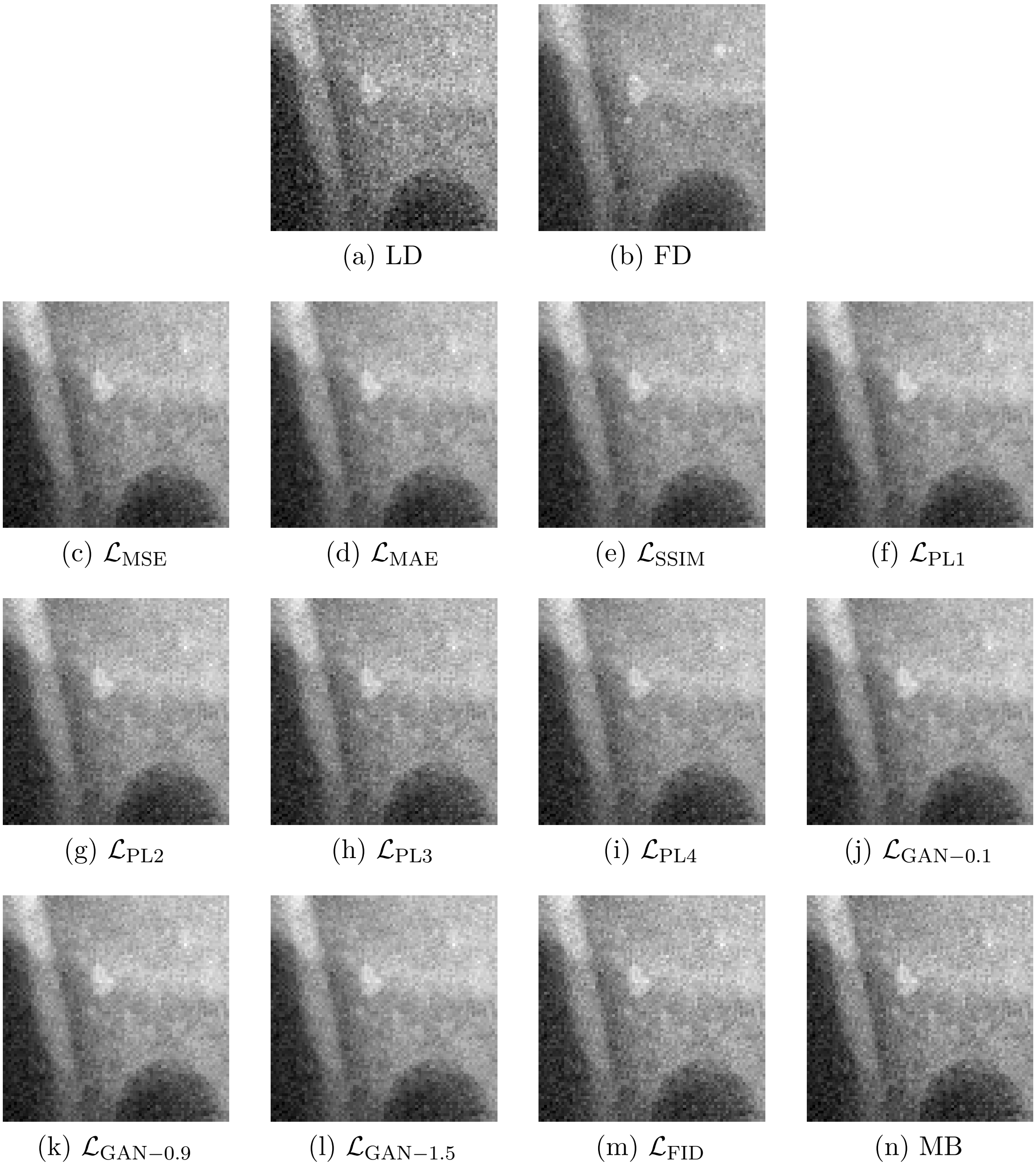
(l) $\mathcal{L}_{\text{GAN}-1.5}$

(m) $\mathcal{L}_{\text{FID}}$

(n) MB

Figure 9 – Illustration of a magnified ROI of an anthropomorphic breast phantom image focusing a simulated cluster of MC. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range.

tested in real cases where the mammography is acquired at a reduced dose, even though it was trained with the simulated injection of noise, since anthropomorphic phantoms are designed to match clinical images as close as possible (CARTON *et al.*, 2011).



(a) LD

(b) FD

(c) $\mathcal{L}_{\text{MSE}}$

(d) $\mathcal{L}_{\text{MAE}}$

(e) $\mathcal{L}_{\text{SSIM}}$

(f) $\mathcal{L}_{\text{PL1}}$

(g) $\mathcal{L}_{\text{PL2}}$

(h) $\mathcal{L}_{\text{PL3}}$

(i) $\mathcal{L}_{\text{PL4}}$

(j) $\mathcal{L}_{\text{GAN}-0.1}$

(k) $\mathcal{L}_{\text{GAN}-0.9}$

(l) $\mathcal{L}_{\text{GAN}-1.5}$

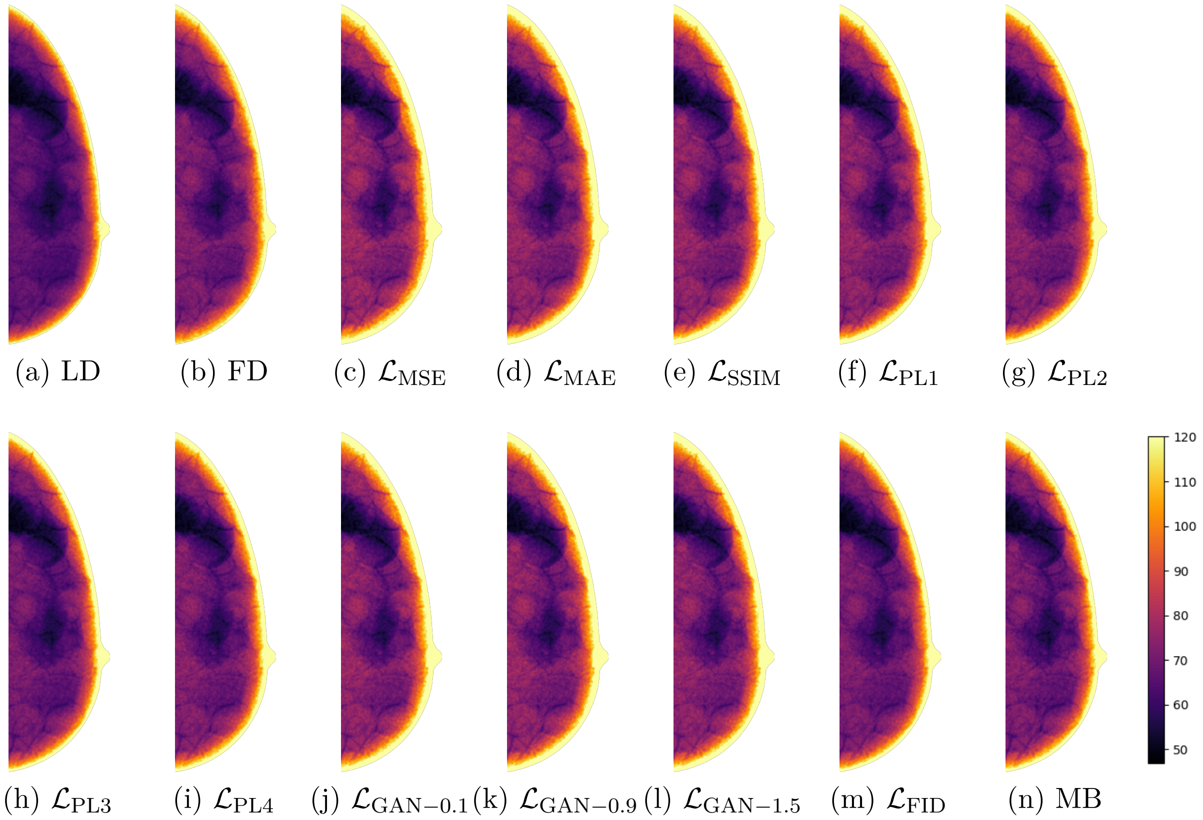
(m) $\mathcal{L}_{\text{FID}}$

(n) MB

Figure 10 – Illustration of a magnified ROI of an anthropomorphic breast phantom image focusing a simulated cluster of MC. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the images were normalized based on the FD and are displayed in the same dynamic range.

2.5.2 Quantitative evaluation

Figs. 11 and 12 illustrate the SNR map inside the breast region of the anthropomorphic breast phantom image. Through a visual inspection, we can see that all the restorations were able to increase the SNR value from the LD. Also, the maps indicate that the restorations with MSE, MAE, SSIM, and adversarial loss achieved the greatest



(a) LD (b) FD (c) $\mathcal{L}_{\text{MSE}}$ (d) $\mathcal{L}_{\text{MAE}}$ (e) $\mathcal{L}_{\text{SSIM}}$ (f) $\mathcal{L}_{\text{PL1}}$ (g) $\mathcal{L}_{\text{PL2}}$

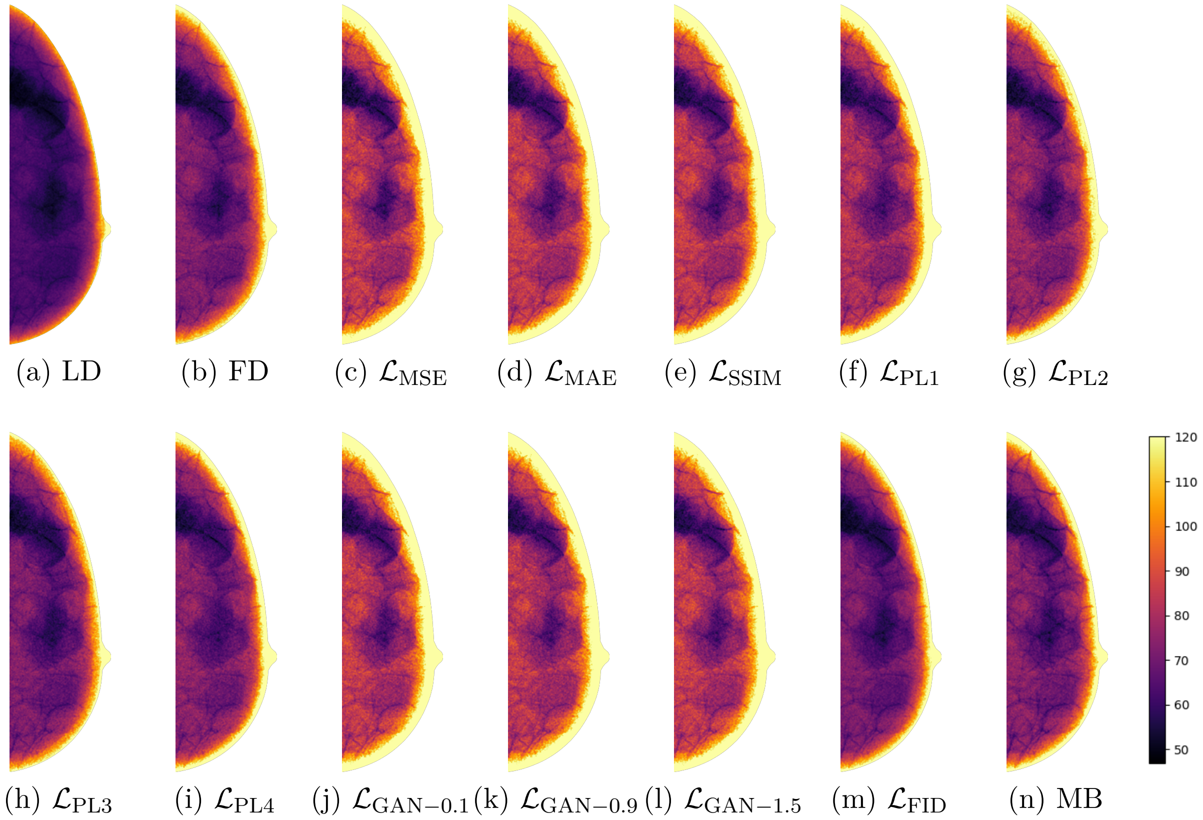
(h) $\mathcal{L}_{\text{PL3}}$ (i) $\mathcal{L}_{\text{PL4}}$ (j) $\mathcal{L}_{\text{GAN}-0.1}$ (k) $\mathcal{L}_{\text{GAN}-0.9}$ (l) $\mathcal{L}_{\text{GAN}-1.5}$ (m) $\mathcal{L}_{\text{FID}}$ (n) MB

Figure 11 – Illustration of the SNR map of the anthropomorphic breast phantom images. (a) LD acquisition for a dose reduction factor of 75%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the maps were adjusted and clipped in the range of 47-120, based on the SNR of the FD image.

values throughout the breast. Table 1 presents the mean SNR value, demonstrating the aforementioned statements. The observance of higher SNR values for these loss functions is explained as they remove more noise compared with the other restorations, thus achieving a lower standard deviation.

From Table 1 it is observable that pixel-wise losses and the adversarial loss yielded the highest SNR, even higher than the SNR of the FD image, thus indicating the best image quality. However, a quick inspection of Fig. 10(d) shows that these loss functions result in some signal smearing/blurring, especially compared to the FD image in Fig. 10(a). This emphasizes the need for a metric sensitive to signal smoothing and residual noise separately. To that end, we adopted the decomposition of the NMSE into $\mathcal{B}^2$ and $\mathcal{R}_{\mathcal{N}}$.

As our primary goal is to restore the LD images to achieve the quality of the standard FD images, we desire a resulting image which has the overall characteristics of the standard FD. Thus, we seek a restoration method that yields an image with similar $\mathcal{R}_{\mathcal{N}}$ compared with the FD and as low $\mathcal{B}^2$ error as possible. This intuition comes from the



(a) LD (b) FD (c) $\mathcal{L}_{\text{MSE}}$ (d) $\mathcal{L}_{\text{MAE}}$ (e) $\mathcal{L}_{\text{SSIM}}$ (f) $\mathcal{L}_{\text{PL1}}$ (g) $\mathcal{L}_{\text{PL2}}$

(h) $\mathcal{L}_{\text{PL3}}$ (i) $\mathcal{L}_{\text{PL4}}$ (j) $\mathcal{L}_{\text{GAN}-0.1}$ (k) $\mathcal{L}_{\text{GAN}-0.9}$ (l) $\mathcal{L}_{\text{GAN}-1.5}$ (m) $\mathcal{L}_{\text{FID}}$ (n) MB

Figure 12 – Illustration of the SNR map of the anthropomorphic breast phantom images. (a) LD acquisition for a dose reduction factor of 50%; (b) FD acquisition; restored images generated by the neural network with the loss function: (c) MSE, (d) MAE, (e) SSIM, (f)-(i) PL1 to PL4, respectively, (j)-(l) adversarial loss with different λ_{adv}, (m) FID and (n) model-based (MB) restoration method. All the maps were adjusted and clipped in the range of 47-120, based on the SNR of the FD image.

fact that our goal is to generate restored images that have similar noise properties to the FD images, and also that we want to keep the underlying signal characteristics as close as the original image, as radiologists tend to dislike overly smoothed images.

To this end, the total MNSE was measured against the pseudo-GT and decomposed into $\mathcal{B}^2$ and $\mathcal{R}_{\mathcal{N}}$ for the FD and for both the radiation dose reduction factors, considering all different loss functions.

Here, we slightly modified how we calculated the MNSE, from Eq. (2.12), so that it could take into account the three folds in the cross-validation and generate enough samples for statistical testing. Instead of computing the mean over the pixels as the first step, as shown in Eq. (2.12), now we calculate the normalized quadratic error (NQE), compute the pixel-wise average over the three folds, and calculated the mean over the pixels as:

$$\text{MNSE} = \frac{1}{\boldsymbol{I}} \sum_{(i,j)\,\in\,\mathcal{I}} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{P} \sum_{p=1}^{P} \text{NQE}. \tag{2.14}$$

For paired t-tests, we obtained the samples after the average over K folds. Note that we

Table 1 – Mean SNR values for the anthropomorphic breast phantom images acquired at the LD (for a dose reduction factors of 75% and 50%) and FD. Results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method.

	75%	50%
LD	69.18	56.41
DL-$\mathcal{L}_{\text{MSE}}$	84.84±0.74	94.78±1.47
DL-$\mathcal{L}_{\text{MAE}}$	85.05±0.48	95.44±0.46
DL-$\mathcal{L}_{\text{SSIM}}$	84.67±0.24	95.51±1.09
DL-$\mathcal{L}_{\text{PL1}}$	83.90±0.19	93.08±0.63
DL-$\mathcal{L}_{\text{PL2}}$	81.56±0.22	86.07±0.29
DL-$\mathcal{L}_{\text{PL3}}$	76.97±0.10	77.73±0.38
DL-$\mathcal{L}_{\text{PL4}}$	79.85±0.11	81.45±0.37
DL-$\mathcal{L}_{\text{GAN}-0.1}$	84.72±0.13	96.39±0.61
DL-$\mathcal{L}_{\text{GAN}-0.9}$	84.91±0.14	95.62±0.35
DL-$\mathcal{L}_{\text{GAN}-1.5}$	85.00±0.38	95.61±0.72
DL-$\mathcal{L}_{\text{FID}}$	77.22±0.66	74.94±0.20
MB	78.39	77.53
FD	78.11	

Table 2 – Quantitative analysis of the total MNSE, decomposed into $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$, for the anthropomorphic breast phantom images acquired at the LD (at a dose reduction factor of 75%) and at the FD. Also, the results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method. Confidence interval is shown for p-value=0.05.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}$(%)	$\mathcal{B}^2$(%)
LD	14.06 [14.04, 14.08]	13.95 [13.93, 13.97]	0.11 [0.10, 0.12]
FD	10.47 [10.46, 10.49]	10.40 [10.38, 10.41]	0.07 [0.07, 0.08]
DL-$\mathcal{L}_{\text{MSE}}$	9.46 [9.45, 9.48]	9.13 [9.11, 9.14]	0.33 [0.32, 0.34]
DL-$\mathcal{L}_{\text{MAE}}$	9.30 [9.29, 9.31]	9.09 [9.07, 9.10]	0.21 [0.20, 0.22]
DL-$\mathcal{L}_{\text{SSIM}}$	9.35 [9.33, 9.36]	9.16 [9.15, 9.17]	**0.19 [0.18, 0.20]**
DL-$\mathcal{L}_{\text{PL1}}$	9.54 [9.52, 9.55]	9.32 [9.31, 9.34]	0.21 [0.20, 0.22]
DL-$\mathcal{L}_{\text{PL2}}$	10.10 [10.08, 10.11]	9.89 [9.87, 9.90]	0.21 [0.20, 0.22]
DL-$\mathcal{L}_{\text{PL3}}$	11.45 [11.43, 11.46]	11.23 [11.21, 11.24]	0.22 [0.21, 0.23]
DL-$\mathcal{L}_{\text{PL4}}$	10.56 [10.55, 10.58]	**10.33 [10.32, 10.35]**	0.23 [0.22, 0.24]
DL-$\mathcal{L}_{\text{GAN}-0.1}$	9.35 [9.34, 9.37]	9.16 [9.15, 9.18]	**0.19 [0.18, 0.20]**
DL-$\mathcal{L}_{\text{GAN}-0.9}$	9.35 [9.34, 9.37]	9.11 [9.10, 9.13]	0.24 [0.23, 0.25]
DL-$\mathcal{L}_{\text{GAN}-1.5}$	9.32 [9.31, 9.34]	9.09 [9.08, 9.10]	0.23 [0.22, 0.24]
DL-$\mathcal{L}_{\text{FID}}$	11.10 [11.09, 11.12]	10.84 [10.83, 10.86]	0.26 [0.25, 0.27]
MB	10.93 [10.92, 10.95]	10.81 [10.80, 10.83]	0.12 [0.11, 0.13]

performed this step for the MNSE and its decomposition. Tables 2 and 3 present the mean values of MNSE results for the dose reduction factors of 75% and 50%, respectively.

Table 3 – Quantitative analysis of the total MNSE, decomposed into $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$, for the anthropomorphic breast phantom images acquired at the LD (at a dose reduction factor of 50%) and at the FD. Also, the results for the restored images generated by neural network with the loss function: MSE, MAE, SSIM, PL1 to PL4, adversarial loss with different λ_{adv}, FID and for the model-based (MB) restoration method. Confidence interval is shown for p-value=0.05.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}$ (%)	$\mathcal{B}^2$(%)
LD	21.79 [21.76, 21.82]	21.64 [21.61, 21.67]	0.15 [0.13, 0.16]
FD	10.47 [10.46, 10.49]	10.40 [10.38, 10.41]	0.07 [0.07, 0.08]
DL-$\mathcal{L}_{\text{MSE}}$	8.27 [8.25, 8.29]	7.54 [7.53, 7.55]	0.73 [0.71, 0.75]
DL-$\mathcal{L}_{\text{MAE}}$	8.16 [8.13, 8.18]	7.48 [7.46, 7.49]	0.68 [0.66, 0.70]
DL-$\mathcal{L}_{\text{SSIM}}$	8.16 [8.14, 8.18]	7.49 [7.48, 7.50]	0.67 [0.65, 0.69]
DL-$\mathcal{L}_{\text{PL1}}$	8.58 [8.56, 8.60]	7.93 [7.92, 7.95]	0.65 [0.63, 0.67]
DL-$\mathcal{L}_{\text{PL2}}$	9.79 [9.77, 9.81]	9.16 [9.14, 9.17]	0.63 [0.62, 0.65]
DL-$\mathcal{L}_{\text{PL3}}$	11.73 [11.71, 11.75]	11.31 [11.30, 11.33]	**0.42 [0.40, 0.43]**
DL-$\mathcal{L}_{\text{PL4}}$	10.94 [10.92, 10.96]	**10.39 [10.37, 10.40]***	0.55 [0.54, 0.57]
DL-$\mathcal{L}_{\text{GAN}-0.1}$	8.05 [8.03, 8.07]	7.38 [7.37, 7.39]	0.67 [0.65, 0.68]
DL-$\mathcal{L}_{\text{GAN}-0.9}$	8.20 [8.18, 8.22]	7.50 [7.49, 7.51]	0.70 [0.69, 0.72]
DL-$\mathcal{L}_{\text{GAN}-1.5}$	8.27 [8.25, 8.28]	7.50 [7.49, 7.51]	0.77 [0.76, 0.78]
DL-$\mathcal{L}_{\text{FID}}$	12.28 [12.26, 12.30]	11.77 [11.75, 11.79]	0.51 [0.50, 0.52]
MB	11.88 [11.86, 11.89]	11.60 [11.59, 11.62]	0.28 [0.26, 0.29]

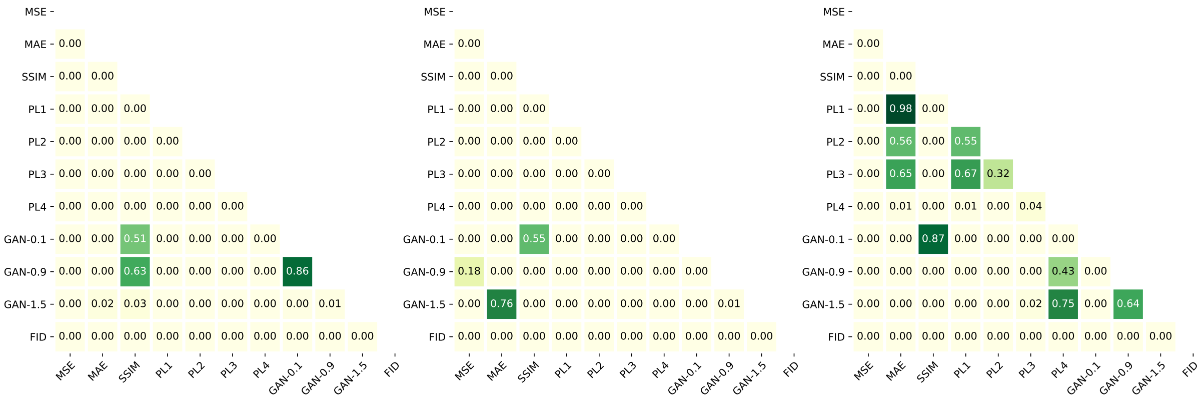
* No statistical difference to the FD (p-values>0.31).

Changing loss functions directly affects the behavior of the deep network in terms of signal preservation and noise suppression. As we can see in Tables 2 and 3, the MSE loss function tends to decrease the $\mathcal{R}_{\mathcal{N}}$ values lower than the standard FD (goal), at the cost of excessively blurring the image, thus increasing the $\mathcal{B}^2$ error. Although this loss function is extensively used in most applications, this blurring behavior is well known in the literature (ZHAO *et al.*, 2016). The MAE loss function (ℓ_1), when compared to the MSE, leads to more image details preservation, observed by lower $\mathcal{B}^2$ values but has a strong denoising effect, noted in the lower $\mathcal{R}_{\mathcal{N}}$ values. For the SSIM loss function case, the model performed better compared to the MSE and MAE, reporting sightly lower $\mathcal{B}^2$ at similar $\mathcal{R}_{\mathcal{N}}$.

The PL function brings an interesting case. There is a tendency to increasing image detail preservation coming from the PL1 to the PL4, where the $\mathcal{R}_{\mathcal{N}}$ slightly increases whereas the $\mathcal{B}^2$ decreases. This behavior is explained by the fact that the deeper neural network layers are responsible for general image characteristics, whereas the initial layers are responsible for local fine details of the image, *i.e.*, primitive information like edges (ERHAN *et al.*, 2009; GU *et al.*, 2018; ZHANG; WU; ZHU, 2018). When looking at the essential properties of the images, in the case of deep layers, the network tends to penalize errors on the underlying signal which causes a decrease in $\mathcal{B}^2$. The opposite happened with the initial layers, where they are trying to match the fine details, thus performing a more aggressive local denoising, decreasing the $\mathcal{R}_\mathcal{N}$ and increasing the $\mathcal{B}^2$ error. The previous discussion brings a consequence that the deeper in the network the loss function analysis is done, less aggressive the denoising and more image details are preserved overall.

FID achieved results close to PL3 and PL4, with higher $\mathcal{B}^2$ in the 75% restoration case. The adversarial loss yields results very similar to the pixel-wise loss functions, *i.e.*, MSE, MAE and SSIM. This is expected since this loss is a combination of an image fidelity loss with an image distribution loss. Looking at the λ_{adv} parameters, increasing its value also increases the mean and standard deviation of $\mathcal{B}^2$. This behavior can be explained by the total loss being more weighed to the adversarial loss, which look at the image as an overall. An extreme case, where just the adversarial loss is used, *i.e.*, $\lambda_{adv} = \infty$ the CNN is very hard to train and could lead to results where the network hallucinate some objects, as known in generative models (SAHARIA *et al.*, 2021). Similar behavior was expected with the PL and FID, since they all use DNN to extract features and optimize the distance between them. However, both FID and PL use pre-trained networks that were trained on very large dataset with natural images. Although pre-trained networks are very good feature extractors, the domain shift from nature images to medical images may cause negative impacts. Therefore, it is suggested to carefully check the quality of restored images in clinical routine.

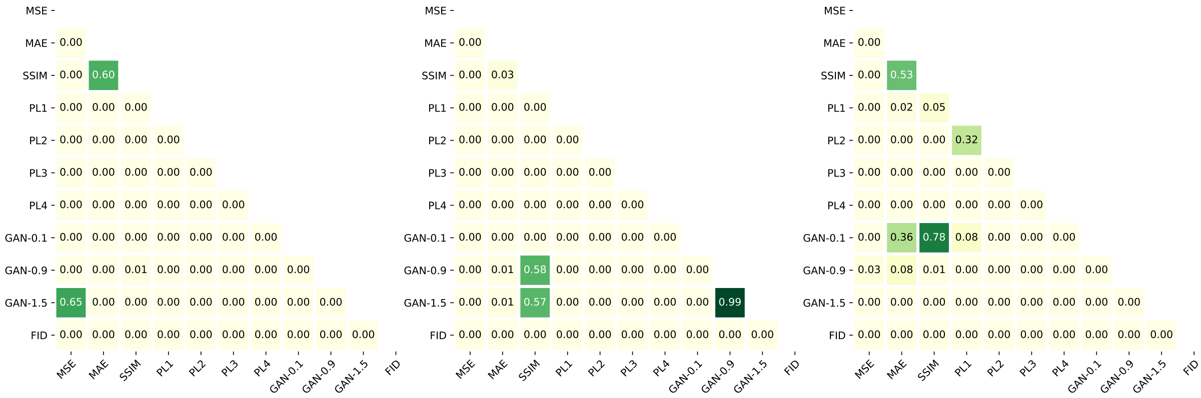
It is important to note the great similarity that the neural networks with PL4 and PL3 have with the FD in terms of $\mathcal{R}_\mathcal{N}$ both for restorations of 75% and 50%. We highlighted in bold the closest $\mathcal{R}_\mathcal{N}$ to the FD and the lowest $\mathcal{B}^2$. Although the $\mathcal{B}^2$ is higher for the deep network compared to the mathematical model, the data-based approach benefits from the fact that it does not need to know any previous information about the equipment and its physics acquisition process.



(a) MNSE (b) $\mathcal{R}_\mathcal{N}$ (c) $\mathcal{B}^2$

Figure 13 – p-value results among all DL losses, for 75% dose reduction, by paired t-test in terms of (a) MNSE, (b) $\mathcal{R}_\mathcal{N}$, and (c) $\mathcal{B}^2$.

Here, we also examine whether any two DL losses have statistically significant differences. The null hypothesis is that the estimated metrics obtained from two losses



(a) MNSE (b) $\mathcal{R}_{\mathcal{N}}$ (c) $\mathcal{B}^2$

Figure 14 – p-value results among all DL losses, for 50% dose reduction, by paired t-test in terms of (a) MNSE, (b) $\mathcal{R}_{\mathcal{N}}$, and (c) $\mathcal{B}^2$.

have identical average. We perform the paired t-test for any two different DL losses and provided the p-values on MNSE and its decomposition in Figs. 13 and 14 for 75% and 50% dose levels respectively. In Fig. 13, when looking at the decomposition of MNSE, PL4 $\mathcal{R}_{\mathcal{N}}$ is statistically different for all other loss functions. Moreover, images restored using the PL4 achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the standard dose; however, it is still statistically different (p-values$<$0.01). For $\mathcal{B}^2$, there is no statistical difference between GAN-0.1 and SSIM and they have the lowest values within all losses. We also notice that many losses have statistically the same $\mathcal{B}^2$. It is interesting to note that even though some losses achieved equal $\mathcal{R}_{\mathcal{N}}$, $\mathcal{B}^2$ was not the same, suggesting the importance of the decomposition of MNSE. The only exception is GAN-0.1 with SSIM which produced equal values for all metrics. In Fig. 14, PL4 is again statistically different from all other losses and achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD. In this case, there is no statistical difference to the FD (p-values$>$0.31). For $\mathcal{B}^2$, PL3 is also statistically different for all other losses and generated the lowest value among them.

Table 4 – Average processing time to run a full restoration on a single raw clinical mammography.

Method	Time (s)
DL-based	9.0
MB	16.5

Finally, Table 4 demonstrates the average time spent by the neural network and also by the MB to restore a single raw clinical image of size 4096 × 3328. Although the deep network takes a very long time for training, for example, roughly 7 hours with MSE and 32 hours with PL4 (respectively the minimum and maximum training time for all loss functions), the restoration process has a processing time of the same order of magnitude as the MB method. Note that we do not intend to compare processing times, as the DL runs on a GPU under Python language while the MB runs on a CPU using a MATLAB code. However, both methods have room for code optimization as fast processing time is especially important for clinical use.

2.6 Conclusion

In this work, we investigated the impact of various loss functions on the quality of LD mammograms restored by deep networks. We used a standard CNN architecture to evaluate such loss functions.

In terms of loss functions, the MSE and MAE had strong denoising properties yielding excessive smoothness in the restored image, while the PL functions preserved images details as we go deeply in the VGG-16 network, *i.e.*, PL3 and PL4 functions preserved more details compared to the PL1. This behavior was observed both in the quantitative results and also in the visual analysis. Furthermore, it is possible to note with both quantitative results and visually the similarity between PL3, PL4 functions and the MB method. The adversarial loss yielded results close to pixel-wise losses, since this function contains one of these losses. It is worth to explore more combinations of λ_{adv} to achieve desired properties and also different training strategies, since GANs are known to be hard to train and susceptible to mode collapse. FID as a loss was demonstrated to be a good alternative as well, achieving visual and objective results close to PLs, MB and FD images.

The fact that we used a physical anthropomorphic breast phantom to validate the proposed methodology reinforces that the neural network is able to restore real LD mammography images. This also implies that the model did not overfit to the training dataset and it is generalizing well for other images with slightly different noise properties.

With this work, we showed the potential of DNNs for digital mammography image restoration and evaluated some well-known loss functions, presenting the strength and weaknesses of each one so we may choose which one is appropriated for each task. Also, with the new training strategy proposed, it is possible to use clinical images and the networks can learn and benefit from a great variability of data and their radiographic factors, as illustrated in Fig. 2.

The limitations of this work are presented as follows. First, since we focused on the comparison among different loss functions, the comparison of different network architectures for the restoration was not considered in this paper. As argued in Shan *et al.* (2018), loss function is relatively more important than network architecture as the loss function has a direct impact on the image quality of the restored images. Second, we did not consider the a combination of different loss functions in this paper, besides the ones used for GAN. Although the combination of several loss functions could improve the results, it will bring extra balancing hyper-parameters to be carefully tuned and exponentially more combinations. We can see this difficulty on choosing λ_{adv} values for GAN. Third, the model was not tested on real patient images as there will be no GT to evaluate the performance of different loss functions. Forth, to achieve a meaningful significant test, we slightly modified the order of mean operations in the MNSE equation so we could get many samples in terms of pixels. In future, one should perform cross-validation with more folds to further analyze the significant tests for each dose level if computational resources permit. Finally, we restricted the training to the dataset representing a local and specific woman population. For future works, real LD mammography images should be tested through the model and cancer detectability evaluations should be performed with radiologists to analyze the relevance of the proposed method in the clinical routine. It is important to test the model with other datasets to see if the model is generalizing well for other populations as the dataset used in the current study is limited to certain characteristics of some population. Also, it is important to evaluate different network architecture and measure the impact of them on the restoration performance.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (62101136), the São Paulo Research Foundation (FAPESP grant 2021/12673-6), the National Council for Scientific and Technological Development (CNPq), the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001), and National Institutes of Health (R01EB026646, R01CA233888, R01CA237267, and R01HL151561).

The authors would like to thank Dr. Andrew D. A. Maidment, from the University of Pennsylvania, for making the anthropomorphic breast phantom images available for this work. The authors also would like to thank the Barretos Cancer Hospital, in particular the medical physicist Renato F. Caron, for providing the clinical images.

3 PAPER 2:ASSESSMENT OF TRAINING STRATEGIES FOR CONVOLUTIONAL NEURAL NETWORK TO RESTORE LOW-DOSE DIGITAL BREAST TOMOSYNTHESIS PROJECTIONS

The material presented in this chapter was published in: Vimieiro, Rodrigo de Barros, Lucas Rodrigues Borges, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Ge Wang, and Marcelo Andrade da Costa Vieira. “Assessment of training strategies for convolutional neural network to restore low-dose digital breast tomosynthesis projections.” In Medical Imaging 2022: Physics of Medical Imaging, vol. 12031, p. 470-479. SPIE, 2022. Available at: https://doi.org/10.1117/12.2609945. The author obtained permission for total and/or partial reproduction of the content from the publisher SPIE.

3.1 Abstract

Convolutional neural networks (CNNs) have been used for image processing tasks such as denoising, deraining, super-resolution and deblurring. It is known that deep models require a large amount of data for the training process. In medical X-ray imaging, data availability is a concern due to patient privacy and radiation exposure. To optimize the training process, here we investigate the effects of training strategies, such as the dataset size, early stop and cross-validation, on the performance of deep neural network. We used the residual network (ResNet) architecture, dedicated to restore low-dose (LD) digital breast tomosynthesis (DBT) raw projections. For this assessment, we generated 500 synthetic breast phantoms through virtual clinical trials software (OpenVCT) and validated them in terms of noise and signal properties. We acquired real raw DBT projections on a commercially available DBT system using a physical anthropomorphic breast phantom and restored the LD projections after training the CNN. Our goal is to restore the LD projections to achieve the same characteristics as the full-dose acquisitions. We found that early stop can be applied in the training process depending on the denoising strength desired for the network. Also, different training realizations are necessary to achieve good results. Furthermore, the training sample size may be smaller compared to other computer vision tasks using deep learning algorithms. Clearly, the amount of data is not the only factor in medical image restoration and other aspects such as network architecture, loss function and data variability must be investigated. The source code developed in this study is available at www.github.com/LAVI-USP/SPIE2022.

3.2 Introduction

In addition to the great success of deep learning (DL) techniques in the computer vision area, DL has been greatly explored in the image processing field (TIAN *et al.*, 2020). Deblurring, deraining, super-resolution and denoising problems have been addressed using neural networks (ARAÚJO; SALVADEO; PAULA, 2021; HAN; BAEK, 2021; PRABHAT *et al.*, 2021; GAO *et al.*, 2020; MARROCCO *et al.*, 2018). However, in areas such as medical imaging, data availability is a concern since DL is known to require a large number of images for the training process (WANG *et al.*, 2018).

Virtual clinical trials (VCTs) have been used to validate medical devices, new techniques, algorithms, and imaging protocols because real clinical trials are time-consuming and expensive. VCTs are capable of simulating different anatomical structures, the physics of the acquisition process and also human interpretation. Using VCT for DL training is a feasible alternative as it is possible to generate as much data as possible with a certain diversity (BARUFALDI *et al.*, 2018). Sahu *et al.* (SAHU *et al.*, 2019) used a VCT software to generate low and high-dose digital breast tomosynthesis (DBT) projections which were used to train a generative adversarial network (GAN) for image denoising. Their network outperformed classical denoising algorithms, such as the block-matching and 3D filtering (BM3D) (DABOV *et al.*, 2007) and K-SVD in terms of the HaarPSI metric. In the work presented by Gao, Fessler and Chan (GAO; FESSLER; CHAN, 2021), virtual phantoms were used as part of the training process of a network designed to denoise DBT slices. They also evaluated the impact of the training dataset size by extracting 400,000 patches from eight physical phantoms and analyzing the contrast to noise ratio (CNR) for 100%, 80%, 65%, 50%, 35% and 20% of the dataset size. Their work reported a large CNR variations for 20% and 35% of the dataset, but some stabilization for greater percentages.

In this work, we investigated different training strategies to restore low-dose DBT projections. The effect of the dataset size in the performance of a deep neural network, was evaluated. Also, approaches such as early stop and cross-validation were also explored. The training dataset was generated using a VCT system. The network was tested using a dataset composed of real DBT projections acquired at a commercially available DBT unit using a physical phantom.

3.3 Materials & methods

3.3.1 Training and testing datasets

Different from object localization and image classification with DL, low-level image processing tasks, such as denoising, rely on learning the relationship between the input and the target in terms of noise and signal characteristics. To investigate the hypothesis that DL approach for image restoration, specifically denoising, does not rely on extremely large datasets as other high-level tasks, we generated training datasets with different sizes using the OpenVCT[1] system developed by the University of Pennsylvania (BAKIC *et al.*, 2017). At the end, we tested the networks using low-dose (LD) DBT projections acquired

[1] www.sourceforge.net/p/openvct

using a physical anthropomorphic breast phantom, at a Selenia Dimensions (Hologic, Inc.) DBT unit, available at the Hospital of the University of Pennsylvania.

3.3.1.1 Training

We used the VCT software to generate the training samples. First, we simulated 500 breast phantoms with different anatomical structures. Then, we configured the software to generate raw DBT projections following the acquisition geometry of the Selenia Dimensions DBT system. To generate pairs of LD and full-dose (FD) projections, we generated noise-free projections from all breast volumes. Then, we inserted quantum and electronic noise using an in-house software package developed by our group in previous work (BORGES *et al.*, 2019). For the FD acquisition, we simulated noise and signal characteristics to achieve signal-to-noise ratio (SNR) equivalent to the standard-dose acquisitions of a real anthropomorphic breast phantom generated using the automatic exposure control (AEC) mode. In addition, we simulated LD acquisitions correspondent to a dose-reduction of 50% of the standard-dose.

When working with virtual images, the image restoration success is dependent on the realism of the simulated images in comparison with the clinical ones. In the work by Bakic *et al.* (BAKIC *et al.*, 2013), the authors validated the virtual breast phantoms simulated with the OpenVCT software through tissue texture analyses. They observed a good similarity with clinical images. In this work, we also validated the noise characteristics of the virtual phantom through SNR and normalized noise power spectrum (NNPS) measurements, as done previously in the work of Borges *et al.* (BORGES *et al.*, 2019). In the breast phantom projections, we measured the point-wise SNR as the mean over the standard deviation (STD) for both real and virtual images. We generated 5 realizations of one virtual phantom so we could measure the point-wise mean and STD. To investigate the noise properties in the frequency domain, we followed the work by Wu, Mainprize and Yaffe (WU; MAINPRIZE; YAFFE, 2012b) to measure the NNPS using a uniform polymethyl methacrylate (PMMA) physical phantom, calculated as follows:

$$NPS(u,v) = \frac{p_s^2}{n_r\, n_p} \sum_{k}^{n_r} \left| \mathcal{FFT}\{h(x,y) \times u_k(x,y)\} \right|^2 \tag{3.1}$$

$$NNPS(u,v) = \frac{NPS}{LAS^2}, \tag{3.2}$$

where (x, y) and (u, v) are 2D spatial coordinates for image and Fourier domain, respectively, p_s is the pixel size in mm, n_r is the number of extracted ROIs, n_p is the ROI total number of pixels, h is a Hanning window, $\mathcal{FFT}$ indicates the discrete Fourier transform and LAS stands for large are signal, which is the mean pixel value squared in the breast region. The

1D plot was calculated as a radially mean of the 2D spectrum. We normalized the NPS through (3.2) by the squared mean pixel values inside the breast.

Our training dataset consists of 256,000 patches, of size 64×64, extracted randomly from the raw DBT projection images of 500 breast phantoms. Indeed, we did not use all the projections since each image generates usually 185 patches. For the training sample size evaluation, we used 100%, 75%, 50% and 25% of the total number of patches. Also, we evaluated an oversized dataset with 1 million patches, approximately 400% of the original dataset. We limited the ROI extraction to the breast tissue region.

3.3.1.2 Testing

Testing samples were generated at the Selenia Dimensions (Hologic, Inc.) DBT system, using the anthropomorphic breast phantom developed at the University of Pennsylvania (CARTON *et al.*, 2011). Such experiments were desirable since it is not feasible to expose patients to X-rays several times (YAFFE; MAINPRIZE, 2011). The testing dataset consists of 20 images at the FD (31kVp and 60mAs) and 10 LD images (31kVp and 30mAs). The FD images were acquired using the AEC mode of the DBT system and the manual mode to yield the LD acquisitions at 50% of the full-dose.

3.3.2 Network and implementation details

We used the commonly known residual network (ResNet) (HE *et al.*, 2016) to perform the image restoration. Besides the popular skip connections, we added a residual layer that goes from the input to the output, as our transformation function is close to the identity matrix. Fig 15 illustrates the utilized network architecture. The CNN was implemented using the PyTorch Deep Learning library[2], trained and tested using an NVIDIA Quadro M5000 GPU with 8GB of RAM. We used Adam optimization (KINGMA; BA, 2014) in all experiments. All implementations details such as learning rate, optimizer parameters and batch size can be found in our code available on GitHub[3].

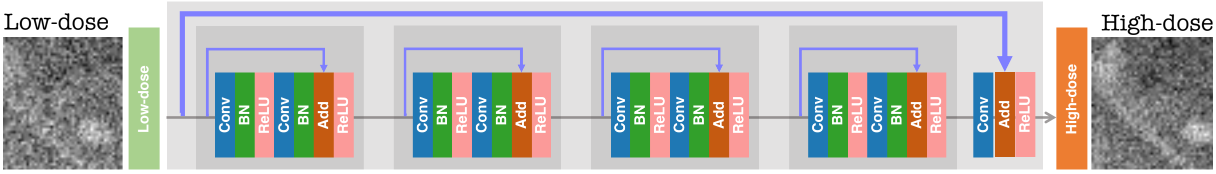


Figure 15 – Illustration of the deep neural network architecture utilized in this work.

We trained the network first with the L1 loss for 60 epochs, so it learns how to reproduce an image without noise. This procedure is similar to the self-supervised learning method presented previously named *Noisier2Noise* (MORAN *et al.*, 2020). In this

[2] www.pytorch.org
[3] www.github.com/LAVI-USP/SPIE2022

work, the authors injected noise in the image to create a pair of images and trained the network towards the noise-free image. Many other self-supervised methods were proposed and can be applied with real clinical data that have no labels (HENDRIKSEN; PELT; BATENBURG, 2020; BATSON; ROYER, 2019; KRULL; BUCHHOLZ; JUG, 2019; NIU *et al.*, 2022). Although true labels are available to us, since we are using simulated images with the VCT software, the adopted methodology is appropriate considering that our goal it to reach the FD image and not a noise-free image. After training with this fidelity loss function, we introduced the perceptual loss (PL) (JOHNSON; ALAHI; FEI-FEI, 2016) also for 60 epochs to retrieve the noise properties of the FD images. In our case, this detail retriever loss function uses the VGG-16 (SIMONYAN; ZISSERMAN, 2014) network pre-trained on the ImageNet dataset. We used the fourth block of this architecture to compute the loss as follow:

$$\mathcal{L}_{\text{PL}} = \frac{1}{w \times h \times c} \sum_{i=1}^{w} \sum_{j=1}^{h} \sum_{k=1}^{c} \left(\Phi(\widehat{x})_{ijk} - \Phi(x)_{ijk}\right)^2 , \tag{3.3}$$

where Φ represents the VGG-16 fourth block output, x the LD input image, $\widehat{x}$ the estimated FD image and $w{\times}h{\times}c$ the patch size. This training approach was also done in our previous work with 2D digital mammography (SHAN *et al.*, 2023). All networks were trained 3 times so that we could compare the results for each realization in terms of objective metrics. For that, we randomly initialized both the weight and bias in the beginning of the training. We performed a testing step for every 10 epochs of the training procedure.

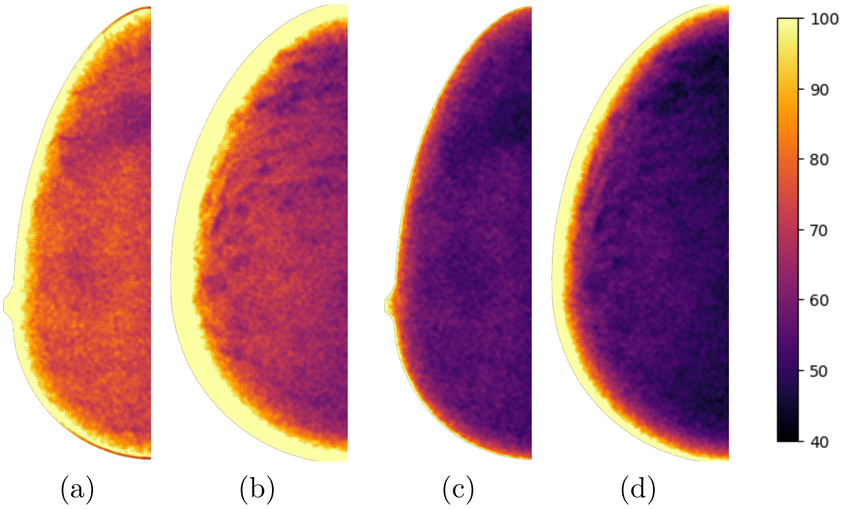


Figure 16 – SNR maps for (a) central DBT projection acquired with the physical anthropomorphic breast phantom at the FD; (b) central projection for the virtual phantom using the OpenVCT software; (c) and (d) are the SNR maps of the same physical and virtual phantoms, respectively, acquired at 50% of the FD.

3.3.3 Objective metrics

To investigate the effectiveness of the restoration method, we used the mean normalized squared error (MNSE) decomposed into two parts: bias squared ($\mathcal{B}^2$) and residual noise ($\mathcal{R}_{\mathcal{N}}$), as in Borges *et al.* (BORGES *et al.*, 2018). In simple words, the metric measures, independently, how a group of images is similar to the ground-truth (GT) in terms of signal and noise. The $\mathcal{B}^2$ value indicates how the underlying signal was impacted in comparison to the GT. In other words, it measures how much blurring and smearing was involved in the denoising process. On the other hand, $\mathcal{R}_{\mathcal{N}}$ measures the noise variance of the restored image. Both are first calculated as a point-wise metric and then the mean value is taken over the breast region of all the projections.

In previous work, our group presented and validated a model-based (MB) denoising pipeline to restore LD DBT projections (BORGES *et al.*, 2017; BORGES *et al.*, 2018). The pipeline uses a variance stabilizing transformation (VST) and the BM3D (DABOV *et al.*, 2007) algorithm to denoise the images. We also measured the MNSE for this method so we could have a benchmark for our networks.

3.4 Results & discussions

In this section, we present and discuss the results related to the restoration of networks trained considering all dataset sizes. First, we illustrate the virtual phantom validation, in terms of noise properties, since the success of the denoising by CNNs is close related to that.

3.4.1 Virtual phantom validation

Figure 16 shows the SNR maps for the central DBT raw projection measured using physical and virtual phantoms at the FD and LD. Although the virtual breast phantoms do not correspond physically to the real one, we can notice that both synthetic and real images approximate well in terms of SNR.

Table 5 – SNR values for the real and virtual phantoms. The third column gives the relative errors from the virtual phantom compared to the real acquisition.

Phantom	SNR	Error
VCT (60mAs)	76.91 ± 19.81	2.29%
VCT (30mAs)	57.77 ± 13.55	0.19%
Physical (60mAs)	78.71 ± 11.53	
Physical (30mAs)	57.66 ± 9.13	

To quantitatively analyze the SNR, we calculated the mean SNR value inside the breast tissue over the 15 projections. Table 5 contains these values for all the images in Fig. 16. We segmented the breast to compute the metric and did not consider the region next to the breast boundaries. The reported error is less than 3%, suggesting that the projections are a good approximation to the real images in terms of signal and noise.

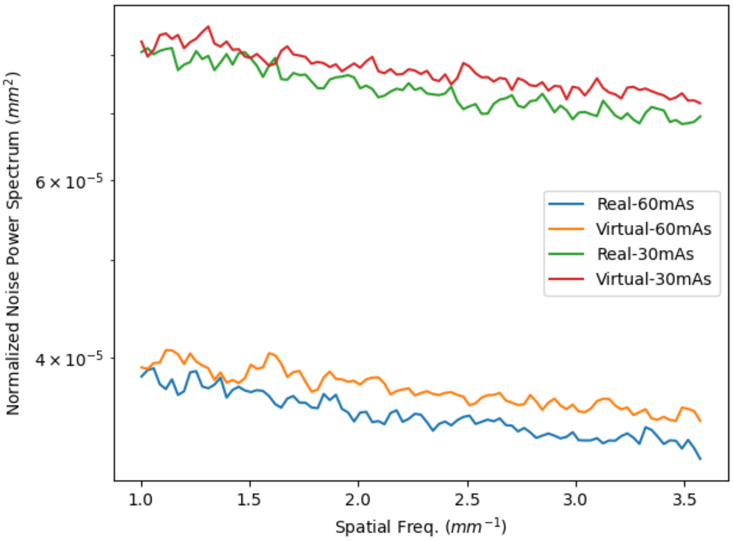


Figure 17 – Normalized noise power spectrum (NNPS) for the physical and virtual PMMA phantoms at the LD and FD.

As different DBT system have distinct method to acquire signals, the noise characteristics is particular for each vendor. As an example, amorphous silicon (a-Si) coupled with a thallium-doped cesium iodide (CsI:Tl) and amorphous selenium (a-Se) detectors commonly report distinct noise properties in the frequency domain (BORGES *et al.*, 2019). Figure 17 illustrates the noise behavior in the frequency domain for both virtual and physical PMMA phantoms. This frequency characterization is important since it validates the estimated noise correlation kernel, which is particular for each DBT system. We can note that the NNPS curves approximates well when comparing real and synthetic images.

3.4.2 Visual analysis

Figure 18 shows regions-of-interest (ROIs) extracted from real central DBT projections acquired with the physical phantom at FD and LD. Also, it shows the results of the restoration for the DL model when trained with different dataset sizes. The MB restoration is illustrated as a benchmark. These ROIs were selected from the realization that resulted in the closest $\mathcal{R}_{\mathcal{N}}$ at the last epoch.

Visually, it is perceptible that the proposed DL approach, trained with 100% of the database, was able to restore the LD acquisitions and generated images with noise characteristics very similar to the FD image. The result for this dataset is also very similar to the MB. For the largest dataset, i.e., 400%, the ROI has less contrast when compared to the 100% and also some noise correlation is visually perceptible. As the amount of training data decreases to 50% and 25% of the database, the restored images gradually

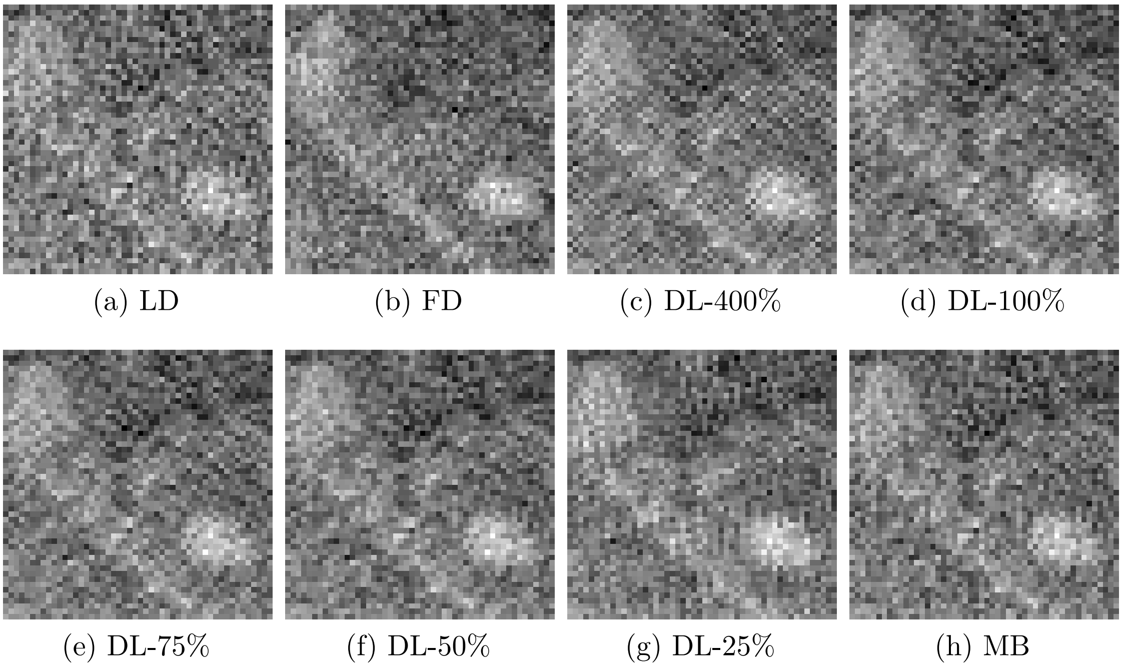


Figure 18 – ROIs extracted from central DBT projections of the physical phantom: (a) LD acquisition; (b) FD acquisition; (c) to (g) DL restoration results for different amounts of training data; (h) restored using the MB pipeline.

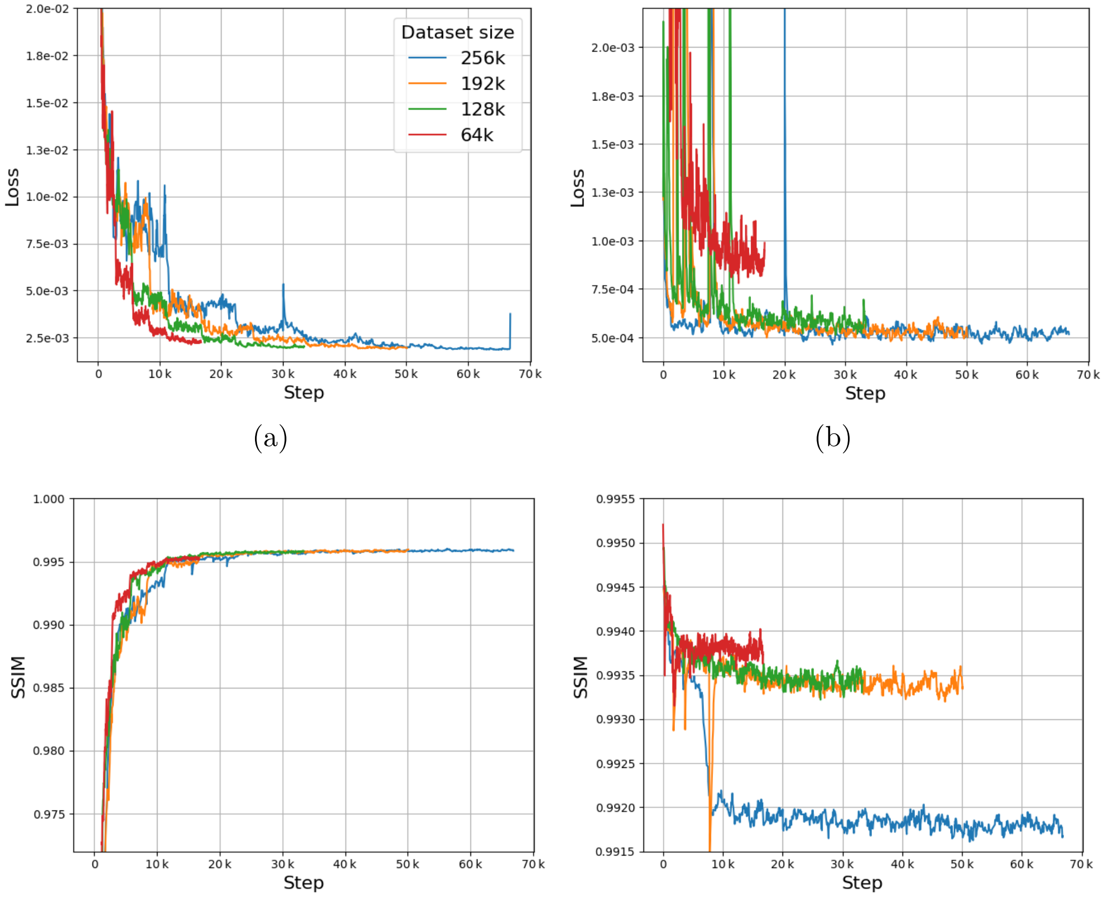
presents image artifacts. There are no visually noticeable differences between the 75% and 100% datasets.

3.4.3 Quantitative evaluation

In this section, we show the objective results in the testing dataset for different sizes of training datasets and also on different epochs of the training stage. Although Fig. 19 shows that all datasets reached close values of loss for both L1 and PL training, their MNSE metrics are distinct. Also, in Fig. 19 we can observe that in the L1 training, the network achieved an SSIM close to 1, measured between the input and the output. However, the PL to retrieve the noise characteristics reduces the SSIM value.

In the MNSE assessment, we evaluated it in five different ways. First, as shown in Tab. 6, we choose within all realizations the results that achieved the closest $\mathcal{R}_{\mathcal{N}}$ regarding the value of the FD acquisition. From this table, we can observe that DL-75% achieved the closest $\mathcal{R}_{\mathcal{N}}$ value. The other training datasets also achieved comparable results, but with higher $\mathcal{B}^2$ value. The largest dataset achieved the lowest $\mathcal{B}^2$ within all networks.

The second evaluation was done concerning the lowest $\mathcal{B}^2$ within the last epoch of all realizations. Table 7 illustrates those results. For this assessment, we can see that the largest dataset achieve the lowest value. Through Tab. 6 and 7 we can observe that different realizations yielded different MNSE values. In some cases, the network achieves good results in terms of $\mathcal{B}^2$ and others in terms of $\mathcal{R}_{\mathcal{N}}$. From these results, we can analyze



(a)

(b)

(c)

(d)

Figure 19 – Illustration of the losses for the (a) L1 and (b) PL for each training step. Also, SSIM calculated at each training step for both (c) L1 and (d) PL losses.

the importance of performing cross-validation also for the assesement of image restoration.

The third and fourth assessments were performed to illustrate the possibility of early stop when training deep CNNs for image restoration depending on the desired task. Analyzing Tab. 8, it is possible to see that, different from Tab. 6, here, the 100% dataset achieved the closest $\mathcal{R}_{\mathcal{N}}$ with only 50 epochs. Nonetheless, we can see that for 75% and 50%, the closest values were achieved at the end of the training, *i.e.*, at 60 epochs. From that, we can observe that the network slowly converges to the $\mathcal{R}_{\mathcal{N}}$ of the FD through the optimization of the PL.

From Tab. 9, we can clearly see that an early stop can be made when low $\mathcal{B}^2$ values are desirable. None of the networks had to train up to the end to achieve the lowest $\mathcal{B}^2$. For 400% and 75%, only 10 epochs were needed. Perhaps with fewer epochs, the network could achieve even lower $\mathcal{B}^2$ values. For the sake of simplicity, we only evaluated the MNSE at every 10 epochs. This table shows an interesting case, where DL-75%, with approximately one-fifth of the data, could achieve a $\mathcal{B}^2$ value very close to the largest dataset.

The fifth assessment was performed because some networks could not deal effectively

Table 6 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R_N}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the closest $\mathcal{R_N}$ to the FD on the last epoch.

	Total MNSE(%)	$\mathcal{R_N}$(%)	$\mathcal{B}^2$(%)	Epoch#	Rlz#
LD (30mAs)	25.02 [24.90, 25.13]	24.86 [24.86, 24.87]	0.15 [0.15, 0.16]	-	-
FD (60mAs)	11.95 [11.71, 12.20]	11.86 [11.86, 11.87]	0.09 [0.09, 0.09]	-	-
DL-400%	13.76 [13.69, 13.82]	12.69 [12.69, 12.70]	1.06 [1.05, 1.07]	60	1
DL-100%	14.66 [14.60, 14.73]	11.74 [11.74, 11.75]	2.92 [2.89, 2.94]	60	1
DL-75%	13.07 [13.04, 13.10]	**11.90** [11.90, 11.90]	1.17 [1.16, 1.18]	60	3
DL-50%	16.63 [16.50, 16.76]	11.60 [11.59, 11.60]	5.03 [4.97, 5.09]	60	3
DL-25%	45.94 [45.73, 46.15]	12.19 [12.18, 12.19]	33.75 [33.43, 34.08]	60	1
MB	13.50 [13.44, 13.55]	13.23 [13.23, 13.24]	0.26 [0.26, 0.27]	-	-

Table 7 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R_N}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on the last epoch.

	Total MNSE(%)	$\mathcal{R_N}$(%)	$\mathcal{B}^2$(%)	Epoch#	Rlz#
LD (30mAs)	25.02 [24.90, 25.13]	24.86 [24.86, 24.87]	0.15 [0.15, 0.16]	-	-
FD (60mAs)	11.95 [11.71, 12.20]	11.86 [11.86, 11.87]	0.09 [0.09, 0.09]	-	-
DL-400%	13.76 [13.69, 13.82]	12.69 [12.69, 12.70]	**1.06** [1.05, 1.07]	60	1
DL-100%	14.67 [14.61, 14.74]	12.97 [12.96, 12.97]	1.71 [1.69, 1.72]	60	2
DL-75%	13.07 [13.04, 13.10]	11.90 [11.90, 11.90]	1.17 [1.16, 1.18]	60	3
DL-50%	12.41 [12.36, 12.45]	9.63 [9.63, 9.63]	2.78 [2.75, 2.81]	60	2
DL-25%	16.88 [16.84, 16.93]	9.03 [9.02, 9.03]	7.86 [7.75, 7.96]	60	3
MB	13.50 [13.44, 13.55]	13.23 [13.23, 13.24]	0.26 [0.26, 0.27]	-	-

with the breast regions next to the skin. So we used a morphological operation to perform an erosion on the breast mask and remove the region close to the borders. Then, we re-calculated the objective metrics for the same realizations of Tab. 7 with this new breast mask. Table 10 illustrates these results. We can observe that in general, the $\mathcal{B}^2$ decreased significantly, confirming that the networks do not effectively treat the boarders very well. Also, the dataset with the lowest $\mathcal{B}^2$ changed. Instead of the largest dataset, now the DL-100% achieved the lowest $\mathcal{B}^2$ value. Moreover, now all DL restorations could achieve values closer to the MB. In this table, we can clearly see that the $\mathcal{B}^2$ slightly increased as the dataset size decreased. Indeed, these results are in conformity with the visual analysis done in Fig. 18, where the largest dataset was not very effectively, and the DL-100% had the best result. The MNSE results for all realizations, dataset sizes and epochs are available on the GitHub web page.

Table 8 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the closest $\mathcal{R}_{\mathcal{N}}$ to the FD one on all evaluated epochs.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}$(%)	$\mathcal{B}^2$(%)	Epoch#	Rlz#
LD (30mAs)	25.02 [24.90, 25.13]	24.86 [24.86, 24.87]	0.15 [0.15, 0.16]	-	-
FD (60mAs)	11.95 [11.71, 12.20]	11.86 [11.86, 11.87]	0.09 [0.09, 0.09]	-	-
DL-400%	12.75 [12.69, 12.81]	11.75 [11.74, 11.75]	1.01 [1.00, 1.02]	10	1
DL-100%	14.44 [14.39, 14.49]	**11.87** [11.87, 11.87]	2.57 [2.54, 2.59]	50	1
DL-75%	13.07 [13.04, 13.10]	11.90 [11.90, 11.90]	1.17 [1.16, 1.18]	60	3
DL-50%	16.63 [16.50, 16.76]	11.60 [11.59, 11.60]	5.03 [4.97, 5.09]	60	3
DL-25%	43.95 [43.76, 44.13]	11.93 [11.92, 11.94]	32.01 [31.69, 32.34]	20	1
MB	13.50 [13.44, 13.55]	13.23 [13.23, 13.24]	0.26 [0.26, 0.27]	-	-

Table 9 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on all evaluated epochs.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}$(%)	$\mathcal{B}^2$(%)	Epoch#	Rlz#
LD (30mAs)	25.02 [24.90, 25.13]	24.86 [24.86, 24.87]	0.15 [0.15, 0.16]	-	-
FD (60mAs)	11.95 [11.71, 12.20]	11.86 [11.86, 11.87]	0.09 [0.09, 0.09]	-	-
DL-400%	12.75 [12.69, 12.81]	11.75 [11.74, 11.75]	**1.01** [1.00, 1.02]	10	1
DL-100%	14.69 [14.61, 14.76]	13.20 [13.19, 13.20]	1.49 [1.47, 1.51]	30	3
DL-75%	11.61 [11.58, 11.64]	10.52 [10.52, 10.53]	1.09 [1.08, 1.09]	10	3
DL-50%	12.23 [12.18, 12.28]	9.59 [9.59, 9.59]	2.64 [2.61, 2.67]	50	2
DL-25%	16.64 [16.60, 16.67]	8.92 [8.92, 8.92]	7.72 [7.62, 7.82]	50	3
MB	13.50 [13.44, 13.55]	13.23 [13.23, 13.24]	0.26 [0.26, 0.27]	-	-

3.5 Conclusions

In this work, we analyzed different training techniques to improve the effectiveness of DBT restoration through CNNs. First, we observed that different network training realizations lead to distinct results in terms of objective metrics. Depending on the task, the network can achieve good results to match the residual noise of the target or to achieve low bias values. It is still not clear which approach is better when it comes to diagnoses by radiologists. Second, an early stop can be done especially in the cases when low bias values are desired. Third, we could observe that the largest dataset, with 1 million patches, did not improve the results overall, achieving lower performance in some cases. This suggests that the amount of data is not the most important factor in medical image restoration, such as in other areas, where it plays an important role. The training sample size may be smaller compared to that needed for other computer vision tasks using DL algorithms. Other aspects such as loss functions, network architecture, diversity of training dataset and training strategies may be as important as the dataset size. In fact, this works has its

Table 10 – Total MNSE (%) and its decomposition on residual noise ($\mathcal{R}_{\mathcal{N}}$) and squared bias ($\mathcal{B}^2$) for DBT projections acquired with the physical anthropomorphic breast phantom at LD and FD in addition to the restoration results using the DL model trained with different amounts of data. The MB results are illustrated as a benchmark. The confidence interval is displayed inside the brackets. The results are regarding the network realization that achieved the lowest $\mathcal{B}^2$ on the last epoch and were calculated using the new breast mask.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}$(%)	$\mathcal{B}^2$(%)	Epoch#	Rlz#
LD (30mAs)	24.87 [24.76, 24.98]	24.83 [24.82, 24.84]	0.04 [0.04, 0.05]	-	-
FD (60mAs)	11.71 [11.66, 11.77]	11.70 [11.69, 11.70]	0.02 [0.01, 0.02]	-	-
DL-400%	12.73 [12.68, 12.79]	12.38 [12.38, 12.39]	0.35 [0.35, 0.35]	60	1
DL-100%	12.96 [12.91, 13.01]	12.69 [12.69, 12.70]	**0.27** [0.27, 0.27]	60	2
DL-75%	11.76 [11.71, 11.80]	11.40 [11.39, 11.40]	0.36 [0.36, 0.36]	60	3
DL-50%	9.62 [9.59, 9.66]	9.17 [9.16, 9.17]	0.46 [0.45, 0.46]	60	2
DL-25%	9.04 [9.01, 9.07]	8.50 [8.49, 8.50]	0.55 [0.54, 0.55]	60	3
MB	13.25 [13.20, 13.31]	13.11 [13.11, 13.12]	0.14 [0.14, 0.14]	-	-

limitations. The VCT software has its limitation for simulating real data. On the other hand, the results indicate that the VCT is a good alternative to train such models, as it achieved results very close to the networks trained on real clinical images. To further improve the results, a fine-tuning method could be done with real images, such as the ones we used in the testing dataset or real clinical images. Another limitation of this work is that we did not test different network architectures and losses.

Acknowledgments

This work was supported in part by the National Council for Scientific and Technological Development (CNPq), by the São Paulo Research Foundation (FAPESP grant 2016/25750-0 and 2018/19888-5), by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001) and National Institute of Health (R01EB026646, R01CA233888, R01CA237267, and R01HL151561).

4 PAPER 3:IMPOSING NOISE CORRELATION FIDELITY ON DIGITAL BREAST TOMOSYNTHESIS RESTORATION THROUGH DEEP LEARNING TECHNIQUES

The material presented in this chapter was published in: Vimieiro, Rodrigo de Barros, Lucas Rodrigues Borges, Ge Wang, and Marcelo Andrade da Costa Vieira. “Imposing noise correlation fidelity on digital breast tomosynthesis restoration through deep learning techniques.” In 16th International Workshop on Breast Imaging (IWBI2022), vol. 122861, p. 13. SPIE, 2022. Available at: https://doi.org/10.1117/12.2626634. The author obtained permission for total and/or partial reproduction of the content from the publisher SPIE.

4.1 Abstract

Digital breast tomosynthesis (DBT) is an important imaging modality for breast cancer screening. The morphology of breast masses and the shape of the microcalcifications are important factors to detect and determine the malignancy of breast cancer. Recently, convolutional neural networks (CNNs) have been used for denoising in medical imaging and have shown potential to improve the performance of radiologists. However, they can impose noise spatial correlation in the restoration process. Noise correlation can negatively impact radiologists’ performance, creating image signals that can resemble breast lesions. In this work, we propose a deep CNN that restores low-dose DBT projections by partially filtering out the noise, but imposes fidelity of the noise correlation between the original and restored images, avoiding artifacts that may resemble signs of breast cancer. The combination of a loss function that calculates the difference in the power spectra (PS) of the input and output images and another one that seeks image visual perception is proposed. We compared the performance of the proposed neural network with traditional denoising methods that do not consider the noise correlation in the restoration process and found superior results in terms of PS for our approach. Our code is available at www.github.com/LAVI-USP/IWBI2022-PSloss.

4.2 Introduction

Digital breast tomosynthesis (DBT) is a powerful imaging modality to help radiologists on early detection of breast cancer on screening programs. Signs of breast cancer are generally related to four mammography findings: morphological characteristics of the tumor mass, cluster of microcalcifications, architectural distortions and breast asymmetry (YAFFE, 2000). Several studies have been done to investigate how the image quality impacts the detection of breast cancer by radiologists (JR *et al.*, 2007; CHAN *et al.*, 2020; BOITA *et al.*, 2021). Image quality issues are commonly associated with spatial resolution, contrast, pixel cross-talk, image blur and noise (BOITA *et al.*, 2021). Methods to improve the quality of an image that has been degraded by one of the aforementioned issues have been extensively investigated. Some studies have investigated the use of model-based (MB) methods in the field of image reconstruction (ZHENG; FESSLER; CHAN, 2017) and image restoration (BORGES *et al.*, 2018), and, more recently, data-driven methods, with artificial neural networks (GAO *et al.*, 2020; SHAN *et al.*, 2023; VIMIEIRO *et al.*, 2022), which also have been achieving good results.

When it comes to noise spatial correlation, Boita *et al.* (BOITA *et al.*, 2021) argue that this degradation "*have a high impact on the radiologists' perceived ability to interpret mammograms*" and also can "*introduce simulated signals that could resemble calcifications*". Indeed, recent works were proposed to deal with correlated noise to restore or reconstruct images, with MB approaches (ZHENG; FESSLER; CHAN, 2017; MÄKINEN; AZZARI; FOI, 2020). It is known that denoising methods can introduce noise correlation in the restored image, especially for convolutional neural networks (CNNs). The objective of this work is to propose a deep CNN that restores low-dose (LD) DBT projections and preserves the noise characteristics in terms of noise correlation, avoiding artifacts that may resemble signs of breast cancer.

4.3 Materials & methods

4.3.1 Training and testing datasets

4.3.1.1 Training

Our training dataset consists of 1,982 retrospective clinical DBT exams, with a total of 29,730 projections, acquired at the Institute of Radiology (InRad), Faculty of Medicine, University of São Paulo (Brazil) under IRB approval[1]. We used a Hologic Selenia Dimensions DBT system (Hologic, Bedford, MA). All the images were used as raw data, *i.e.*, DICOM "for processing". All clinical images were carefully anonymized to preserve patients' medical records. The method proposed by Borges *et al.* (BORGES *et al.*, 2017) was used to inject quantum and electronic noise on the projections to simulate LD acquisitions of 50% of the standard radiation dose (VIMIEIRO *et al.*, 2022).

4.3.1.2 Testing

We tested the proposed neural networks with an anthropomorphic physical breast phantom designed by the University of Pennsylvania (CARTON *et al.*, 2011) and produced by CIRS, Inc. (Reston, VA). The phantom was exposed 20 times with radiographic factors (31 kVp and 60 mAs) determined by automatic exposure control (AEC). We then halved the mAs value in manual mode (30mAs) and acquired 10 exposures, at the same position, to obtain samples with a 50% radiation dose reduction rate. For image quality assessment, we used the mean normalized squared error (MNSE), decomposed into residual noise

[1] CAAE #56699016.7.0000.0065

($\mathcal{R}_{\mathcal{N}}$) and bias squared ($\mathcal{B}^2$). This decomposition allows the evaluation of the restoration methods in terms of signal smoothing and residual noise separately. This metric was proposed by Borges *et al.* (BORGES *et al.*, 2018) and further information is available in Shan *et al.* (SHAN *et al.*, 2023). We also measured the power spectrum (PS) of the restored images (WU; MAINPRIZE; YAFFE, 2012b; KAVURI; DAS, 2020).

4.3.2 Network and implementation details

We used a new enhanced ResNet architecture, with hierarchical skip connections, proposed by Shan *et al.* (SHAN *et al.*, 2023), for all training schemes. The CNN was implemented using the PyTorch Deep Learning library[2], trained and tested using an NVIDIA TITAN Xp with 12GB of RAM. Overall, five networks were trained to compare their performance in terms of objective metrics. We first pre-trained the networks with an L1 loss function for 60 epochs. We halved the initial learning rate (LR) each 10 epochs. After that, we set the LR to 100 times smaller, *i.e.*, $1e^{-5}$, and trained the networks with different strategies for a single epoch (1880 steps). We considered three different strategies to impose noise correlation fidelity on the network output. First, we adopted the mean normalized absolute error (MNAE) between the 2D PS of the input and the output. Second, we considered the MNAE between the PS 1D. Third, we made a combination of the perceptual loss (PL) (JOHNSON; ALAHI; FEI-FEI, 2016) and the PS 2D. For this third method, we assessed different weight values to scale the PS loss contribution, with the following equation:

$$\mathcal{L}_{PL+PS} = \mathcal{L}_{PL4} + \lambda * \mathcal{L}_{PS2D}. \tag{4.1}$$

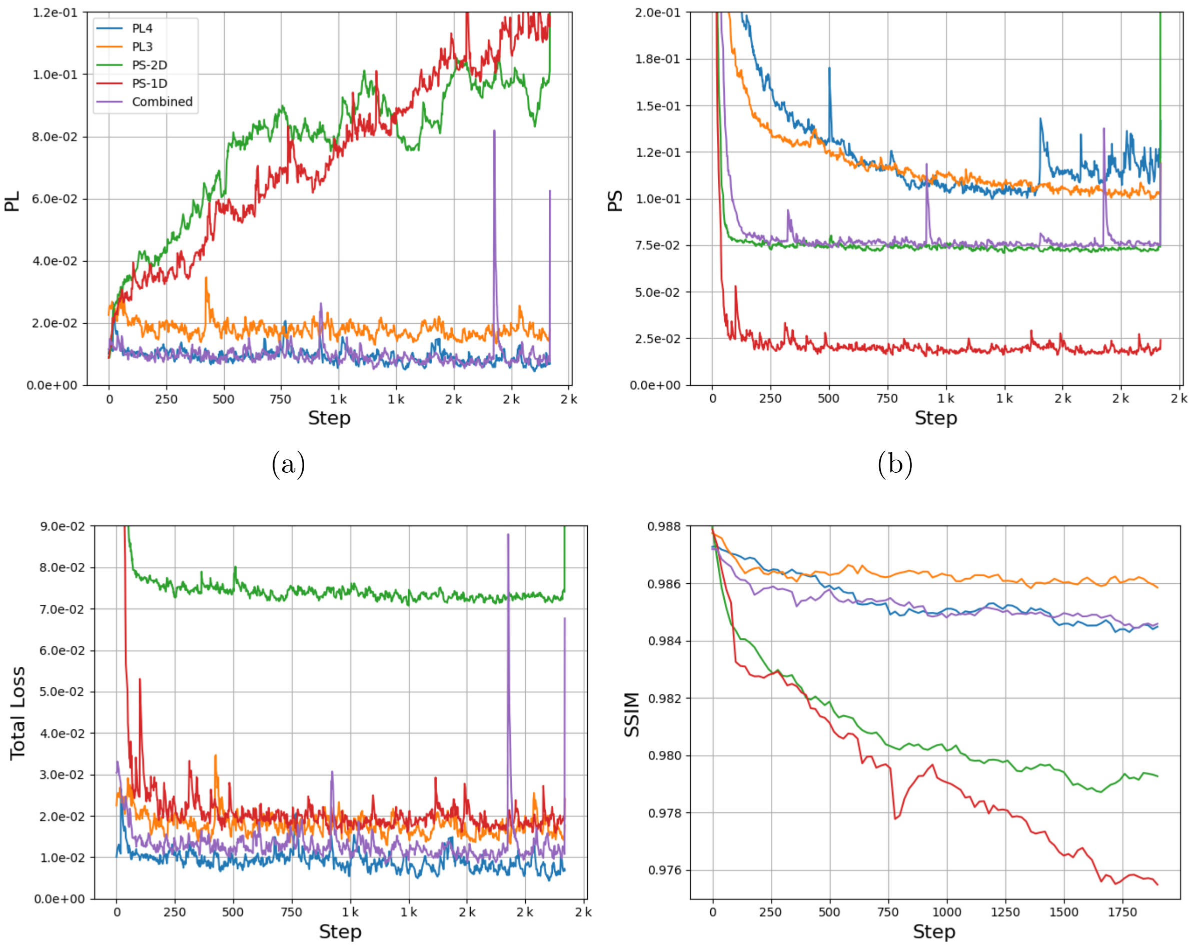
These values were sampled from a uniform distribution. Also, we trained two different networks, with PL on the third and fourth block, *i.e.*, PL3 and PL4 respectively, for comparison. During training we evaluated the structural similarity index (SSIM) (WANG *et al.*, 2004) and the PS. We used 256,000 patches, randomly extracted from the training dataset. Patches that were outside the breast region and also next to the breast skin were removed. This step was done to avoid abrupt changes in the PS and to facilitate the training procedure. For cross-validation, we performed the network fine-tuning 10 times. We used the model-based (MB) restoration, proposed by Borges *et al.* (BORGES *et al.*, 2018), as a reference for the proposed networks.

4.4 Results & discussions

Figure 20 illustrates the PL, PS loss, total loss and SSIM during the training procedure of one realization, for all the proposed networks. In the PL case, we can see that PL3 and PL4 converge and stabilize very fast. However, for the PS networks, the perceptual metric increases along with the training (Fig. 20 (a)). For the PS loss, we can

[2] www.pytorch.org

observe that although the PL networks are not directly optimizing the PS loss, they slowly decrease the PS loss value. This is explained by the fact that the PL is approximating the $\mathcal{R}_{\mathcal{N}}$ of the input and the output, but not considering the noise correlation itself. On the other hand, the PS networks and the network with the combined loss decreased and stabilized the PS values faster than the PL networks (Fig. 20 (b)). The combined loss benefits from both characteristics, *i.e.*, it optimizes the PS while keeping the PL values. Fig. 20 (c) illustrates the total loss, *i.e.*, the combination of the PL and the PS loss. In the cases where either the PL or the PS is optimized, the total loss would be the same. Only in the case of the combination of both of them, the total loss result will be different. We can see in this figure that all losses decrease as the network is trained. Fig. 20 (d) illustrates the SSIM measurement between the input and the target along with the training for all losses. This graphic illustrates a similar behavior as the PL, where PS networks tend to decrease the SSIM value along with the training steps. On the other hand, the PL networks and the combined loss slightly decrease the SSIM values compared to the PS ones.



(a)

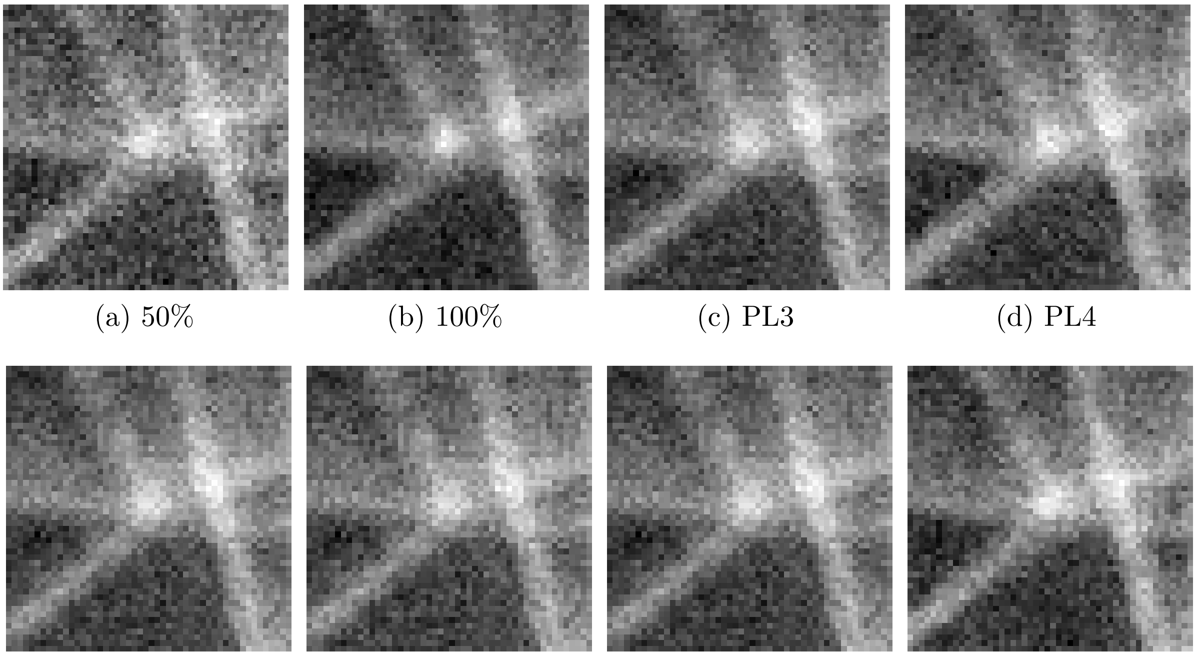
(b)

(c)

(d)

Figure 20 – Illustration of (a) PL, (b) PS loss, (c) total loss and (d) SSIM on the training stage of all the proposed neural networks.

Figure 21 shows regions of interest (ROIs) extracted from the central projection of DBT acquisitions of the physical breast phantom at LD, FD, and the restored images generated from all the networks and from the MB restoration method. In general, visually, all networks were able to successfully restore LD DBT acquisitions, achieving similar results to the MB.



(a) 50% (b) 100% (c) PL3 (d) PL4

(e) PS-1D (f) PS-2D (g) Combined Loss (h) MB

Figure 21 – ROIs extracted from the central projection of DBT acquisitions of the physical breast phantom at (a) LD; (b) FD; CNN output with (c) PL3, (d) PL4, (e) PS-1D, (f) PS-2D, (g) the combined loss (PL+PS) and (h) the Model-based (MB) restoration.

Table 11 shows the MNAE values from each PS curve to the FD one, *i.e.*, the distance between the curves. As discussed before, we evaluated several weights in the combined loss function to tune this hyper-parameter. Within all realizations, we show the lowest value for each restoration method. In bold, we can see that the combined loss with $\lambda = 0.037$ has the lowest MNAE value, *i.e.*, its PS curve is the closest to the FD one. Then, throughout the paper we restricted all the results regarding the combined loss to this specific weight.

Figure 22 illustrates the values from Tab. 11, where $\lambda = 0$ means only PL4 and $\lambda = 1$ PS-2D loss only. We can see that the chosen value ($\lambda = 0.037$) achieved lower MNAE compared to the PL4 itself.

Figure 23 (a) illustrates the PS measurements for the DBT central projection acquired with the anthropomorphic breast phantom at the testing stage. According to Kavuri and Das (KAVURI; DAS, 2020), anatomical noise dominates in the PS on frequencies bellow 1 mm^{-1}, while quantum noise prevails on higher frequencies. Then, Fig. 23 (b) zooms the PS region above 1 mm^{-1} for better visualization. These curves

Table 11 – MNAE metric between the PS curves (Fig. 24 (b)) of different DL losses and the FD acquisition. Hyper-parameter tuning of the combined loss function weight.

Loss	MNAE	Rlz#
LD	0.775	
PL3	0.060	6
PL4	0.057	1
Combined-0.003	0.041	7
Combined-0.023	0.031	6
Combined-0.027	0.034	1
Combined-0.037	**0.027**	7
Combined-0.044	0.032	3
Combined-0.143	0.050	6
Combined-0.212	0.072	10
Combined-0.375	0.096	2
Combined-0.624	0.131	10
Combined-0.939	0.101	5
PS-2D	0.109	10
PS-1D	0.051	5
MB	0.038	

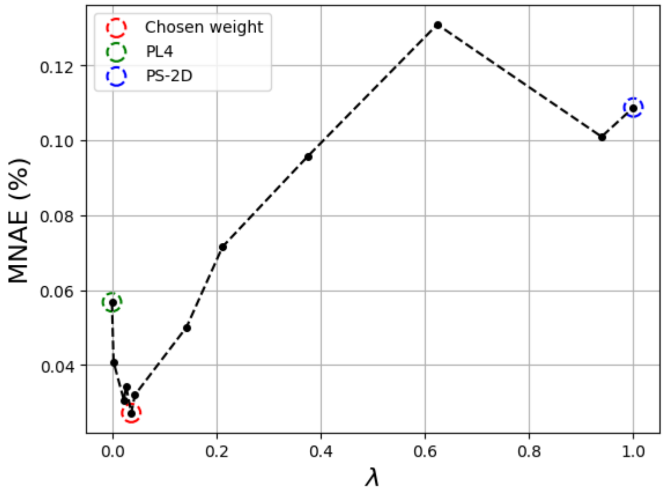
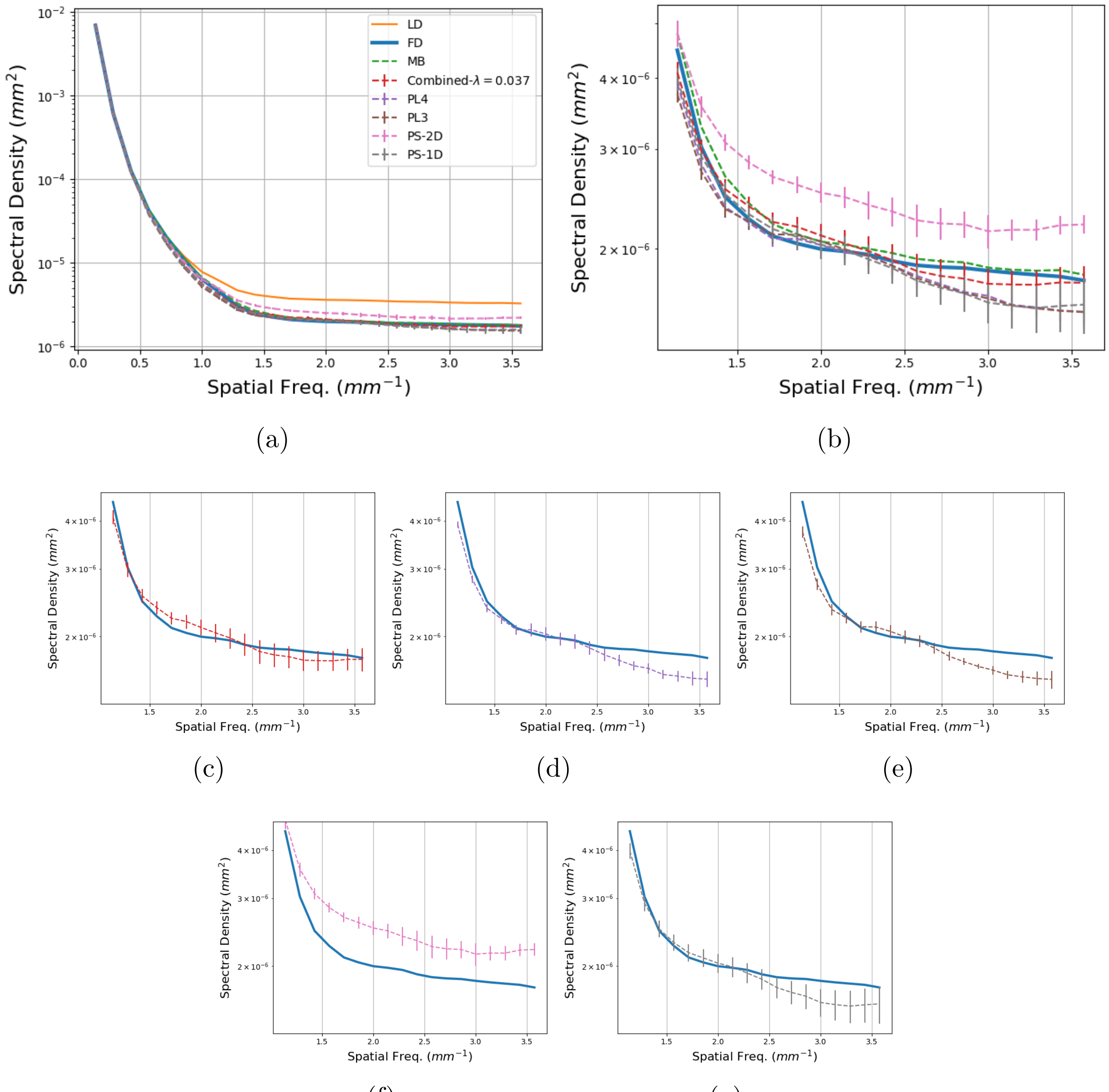


Figure 22 – Visual representation of the relation between the weighting factor (λ) and the MNAE metric between the PS curves.

represent the mean PS measurement with the respective standard deviation (SD) for all training realizations, *i.e.*, the cross-validation process. We can see through the graphics that optimizing the PS only, either the 1D or the 2D, is an unstable process in the testing phase. The SD for each point is relatively high in comparison with the PLs. For the PL curves, although they have lower SD, for higher frequencies the PS deviate from the FD one. Combining both losses can benefit from the properties of each of them, *i.e.*, constraining the PS while stabilizing the training process. To facilitate the visualization of the SD of each zoomed PS of the DL methods, we plotted them separately from (c) to (g).



(a)

(b)

(c)

(d)

(e)

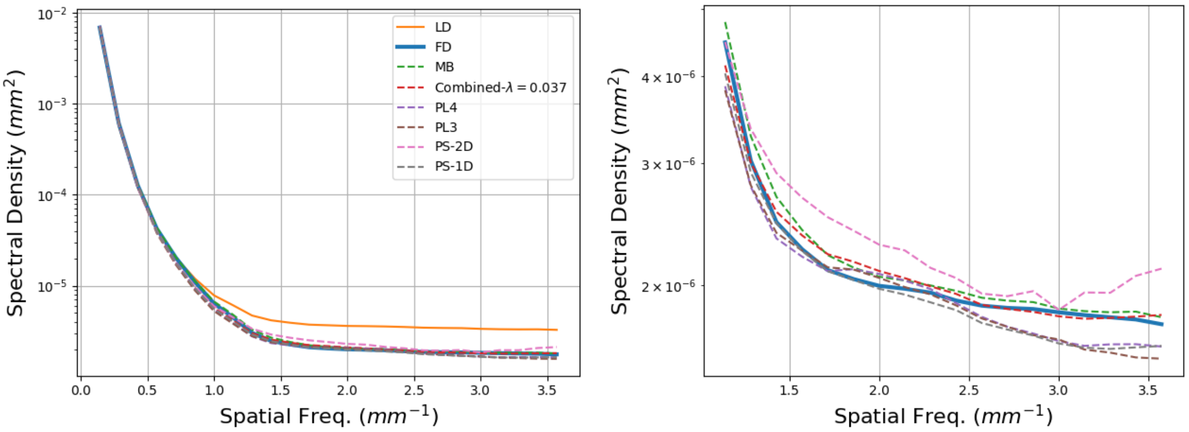
(f)

(g)

Figure 23 – Illustration of (a) the mean PS measurement for the DBT central projection acquired with the physical breast phantom at the testing stage for all training realizations and (b) zoom on PS high frequencies (above 1 mm^{-1}) along with the standard deviation (SD). To facilitate the visualization of the SD of each zoomed PS of the DL methods, we plotted them separately from (c) to (g).

While we can see the behavior of the PS curves throughout different training realizations, it is important to inspect the results of a certain realization. In Fig. 24 we show the PS that achieved the lowest MNAE for each restoration method within all training realizations. We can see in the zoomed PS curve that the combined loss approximates well to the FD one (Fig. 24 (b)).

As our primary goal is to restore the LD DBT projections to achieve the same characteristics of the FD images. Thus, we seek a restoration method that yields an image with similar PS, MNSE and $\mathcal{R}_{\mathcal{N}}$ compared with the FD, with the lowest $\mathcal{B}^2$ error. Table 12



(a) (b)

Figure 24 – Illustration of (a) PS on the physical breast phantom that achieved the lowest MNAE for each restoration method within all training realizations; (b) a zoom on PS high frequencies.

shows the mean value of the MNSE, decomposed into $\mathcal{R}_\mathcal{N}$ and $\mathcal{B}^2$, calculated for the FD and LD acquisition and also for the proposed networks with the loss function PL3, PL4, PS-2D, PS-1D and combined loss (PL4+PS-2D), considering all projections acquired with the physical anthropomorphic breast phantom. The values correspond to the training realization (Rlz#) that achieved the best PS in terms of MNAE, as illustrated on Tab. 11. We also measured the MNSE for the MB as a benchmark.

Table 12 – Mean value of the total MNSE, $\mathcal{R}_\mathcal{N}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the best PS in terms of MNAE, described in the last column. The confidence interval is displayed inside the brackets.

	Total MNSE(%)	$\mathcal{R}_\mathcal{N}$(%)	$\mathcal{B}^2$(%)	Rlz#
LD (30mAs)	24.87 [24.76, 24.98]	24.83 [24.82, 24.84]	0.04 [0.04, 0.05]	
FD (60mAs)	11.71 [11.66, 11.77]	11.70 [11.69, 11.70]	0.02 [0.01, 0.02]	
PL3	12.30 [12.26, 12.35]	11.97 [11.96, 11.97]	0.33 [0.33, 0.34]	6
PL4	12.25 [12.20, 12.29]	**11.93** [11.93, 11.94]	0.32 [0.31, 0.32]	1
PS-2D	15.72 [15.65, 15.79]	15.37 [15.36, 15.37]	0.35 [0.34, 0.35]	10
PS-1D	12.61 [12.56, 12.65]	12.31 [12.30, 12.31]	0.30 [0.29, 0.30]	5
PL4 & PS-2D	13.11 [13.06, 13.16]	12.83 [12.82, 12.83]	**0.28** [0.28, 0.28]	7
MB	13.25 [13.20, 13.31]	13.11 [13.11, 13.12]	0.14 [0.14, 0.14]	

We can observe that optimizing the PS-1D as a loss function is a good surrogate for the PL as they achieved similar results in terms of $\mathcal{B}^2$ and $\mathcal{R}_\mathcal{N}$. As mentioned before, optimizing the PS-2D is not a stable process, as we can also see in the MNSE values, where the $\mathcal{R}_\mathcal{N}$ value is greater than the other methods. For the combined loss, the method achieved better results, with $\mathcal{B}^2$ slightly lower than the PL and slightly higher $\mathcal{R}_\mathcal{N}$, but lower than the MB. Also, there is a trade-off between $\mathcal{B}^2$ and $\mathcal{R}_\mathcal{N}$, where networks with low $\mathcal{B}^2$ values have relatively higher $\mathcal{R}_\mathcal{N}$ and vice-versa.

We show in Tab. 13 and 14 the MNSE values for the realization that achieved the closest $\mathcal{R}_\mathcal{N}$ and the lowest $\mathcal{B}^2$, respectively. Such results illustrate that different training realization result in different images properties. Depending on the application, some realization might be preferable to others. In general, we can realize that the proposed combined loss can achieve good results for all applications. Such cross-validation was presented before by our group on virtual clinical images (VIMIEIRO *et al.*, 2022).

Table 13 – Mean value of the total MNSE, $\mathcal{R}_\mathcal{N}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the closest $\mathcal{R}_\mathcal{N}$ to the FD values, described in the last column. The confidence interval is displayed inside the brackets.

	Total MNSE(%)	$\mathcal{R}_\mathcal{N}$(%)	$\mathcal{B}^2$(%)	Rlz#
LD (30mAs)	24.87 [24.76, 24.98]	24.83 [24.82, 24.84]	0.04 [0.04, 0.05]	
FD (60mAs)	11.71 [11.66, 11.77]	11.70 [11.69, 11.70]	0.02 [0.01, 0.02]	
PL3	12.04 [11.99, 12.08]	**11.67** [11.66, 11.67]	0.37 [0.36, 0.37]	4
PL4	11.97 [11.93, 12.01]	11.65 [11.65, 11.66]	0.32 [0.31, 0.32]	2
PS-2D	15.97 [15.89, 16.05]	15.53 [15.53, 15.54]	0.44 [0.43, 0.44]	1
PS-1D	11.88 [11.83, 11.92]	11.44 [11.43, 11.44]	0.44 [0.44, 0.44]	3
PL4 & PS-2D	12.50 [12.46, 12.55]	12.18 [12.18, 12.19]	0.32 [0.32, 0.33]	2
MB	13.25 [13.20, 13.31]	13.11 [13.11, 13.12]	0.14 [0.14, 0.14]	

Table 14 – Mean value of the total MNSE, $\mathcal{R}_\mathcal{N}$ and $\mathcal{B}^2$ calculated for all DBT projections acquired with the physical breast phantom at the FD and the LD. Also, the results for the restored images generated by the proposed neural networks and the MB method. These results correspond to the realization (Rlz#) that achieved the lowest $\mathcal{B}^2$, described in the last column. The confidence interval is displayed inside the brackets.

	Total MNSE(%)	$\mathcal{R}_\mathcal{N}$(%)	$\mathcal{B}^2$(%)	Rlz#
LD (30mAs)	24.87 [24.76, 24.98]	24.83 [24.82, 24.84]	0.04 [0.04, 0.05]	
FD (60mAs)	11.71 [11.66, 11.77]	11.70 [11.69, 11.70]	0.02 [0.01, 0.02]	
PL3	12.36 [12.31, 12.40]	12.04 [12.04, 12.05]	0.32 [0.31, 0.32]	3
PL4	11.97 [11.93, 12.01]	11.65 [11.65, 11.66]	0.32 [0.31, 0.32]	2
PS-2D	16.57 [16.49, 16.65]	16.39 [16.38, 16.39]	**0.18** [0.18, 0.19]	9
PS-1D	12.21 [12.16, 12.26]	11.93 [11.92, 11.93]	0.28 [0.28, 0.29]	9
PL4 & PS-2D	12.91 [12.87, 12.96]	12.64 [12.64, 12.65]	0.27 [0.27, 0.27]	10
MB	13.25 [13.20, 13.31]	13.11 [13.11, 13.12]	0.14 [0.14, 0.14]	

4.5 Conclusions

In this work, we were able to observe that training CNNs with a combination of two loss functions, one that imposes fidelity in the noise correlation (PS) and another that optimizes the visual perception (PL), can generate restored images that maintain the shape of the PS of the input image. The network trained with the combination of loss functions generated restored images whose PS was very close to the FD, in addition to generating close values of $\mathcal{R}_{\mathcal{N}}$ and $\mathcal{B}^2$. Optimizing only the PS does not guarantee a successful restoration, since in this case high values of $\mathcal{R}_{\mathcal{N}}$ and MNAE were observed for the PS-2D loss function. The PS-1D generated good results in terms of MNSE, but not for the MNAE. Furthermore, the PS-1D loss function could be used as a substitute for PL, due to its ability to preserve image details in the restored image. For future work, the PS-1D can also be used in combined loss, the same way we did for the PS-2D.

Acknowledgments

This work was supported in part by the São Paulo Research Foundation (FAPESP grant 2021/12673-6 and 2022/04646-1), by the National Council for Scientific and Technological Development (CNPq) and by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001). The authors would like to thank Dr. Andrew D. A. Maidment, from the University of Pennsylvania, for making the anthropomorphic breast phantom images available for this work.

5 PAPER 4:LEARNING IN A VARIANCE STABILIZATION DOMAIN TO RESTORE LOW-DOSE DIGITAL BREAST TOMOSYNTHESIS PROJECTIONS

The material presented in this chapter was submitted to *IEEE Transactions on Medical Imaging* journal and it is currently under review. The authors are: Rodrigo de Barros Vimieiro, Chuang Niu, Hongming Shan, Lucas Rodrigues Borges, Ge Wang, and Marcelo Andrade da Costa Vieira.

5.1 Abstract

Digital breast tomosynthesis (DBT) exams should utilize the lowest possible radiation dose while maintaining sufficiently good image quality for accurate medical diagnosis. In this work, we propose a model-based deep-learning framework to restore low-dose (LD) raw DBT projections to achieve an image quality equivalent to a standard full-dose (FD) acquisition. The proposed framework benefits from priors in terms of layers that were inspired by traditional model-based restoration methods. Considering a Poisson-Gaussian noise model, the model is trained in a variance-stabilizing domain, implemented as a layer, where the noise becomes approximately Gaussian, signal-independent, and with unity variance. Additionally, we propose employing a combined bias-residual noise loss function to control the final noise characteristics, potentially enhancing its suitability for clinical applications by allowing the network to be adjusted to attain the desired image quality. The training dataset was composed of clinical data acquired at the standard FD and LD pairs obtained by the injection of quantum noise. The network was tested using real DBT projections acquired with a physical anthropomorphic breast phantom. The proposed framework achieved superior results in terms of mean normalized squared error, structural similarity index, signal-to-noise ratio, and training time compared with models trained with traditional data-driven methods. We observed improvement in restoration for all three architectures that we used, reinforcing the fact that the proposed framework is agnostic to the network architecture. The proposed approach can be extended for other medical imaging application that requires LD acquisitions. The source code is available online.

5.2 Introduction

Digital breast tomosynthesis (DBT) is a pseudo-3D imaging modality in which X-ray projections of the breast are acquired along an arc, allowing the reconstruction of a volume that minimizes tissue overlap and thus improves lesion detection if compared to standard 2D mammography (HOOLEY; DURAND; PHILPOTTS, 2017; LAI *et al.*, 2020).

DBT is emerging as a highly important tool for population-based breast cancer screening programs. Since DBT exams utilize X-rays during image acquisition, the radiation dose must be kept as low as possible, while preserving image quality for clinical utility. In this context, image restoration methods can play a significant role, as they enable the enhancement of image quality while maintaining the radiation dose at an acceptable level for the patient.

The restoration of low-dose (LD) x-ray images is an active research topic and it has been approached using different methods, such as traditional iterative reconstructions (DAS *et al.*, 2010; XU *et al.*, 2015; ZHENG; FESSLER; CHAN, 2017) and denoising techniques (WU; MAINPRIZE; YAFFE, 2012a; BORGES *et al.*, 2017; BORGES *et al.*, 2018). More recently, data-driven methods using neural networks were investigated, achieving promising results (KANG; MIN; YE, 2017; CHEN *et al.*, 2017a; WOLTERINK *et al.*, 2017; CHEN *et al.*, 2017b; KANG *et al.*, 2018; SHAN *et al.*, 2018; SHAN *et al.*, 2019; YIN *et al.*, 2019; WU *et al.*, 2021).

Particularly for DBT, Liu *et al.* (LIU *et al.*, 2018) used two breast specimens to train a convolutional neural network (CNN) for the restoration of DBT images. The LD and full-dose (FD) image pairs were both acquired using a DBT system and the network was trained in the projection domain. Sahu *et al.* (SAHU *et al.*, 2019) trained a generative adversarial network (GAN) to perform denoising using LD and FD pairs of digital phantoms. Gao, Fessler and Chan (GAO; FESSLER; CHAN, 2021) proposed a Wasserstein generative adversarial network (WGAN) for denoising DBT slices. The network was trained on digital and physical phantoms and tested on clinical images. Both mean squared error (MSE) and adversarial loss were used in the training step.

The works mentioned above rely solely on the data available for training and do not include any prior knowledge about signal or noise characteristics. However, recent work showed evidence that adding priors to the CNN architecture may improve performance (WU *et al.*, 2017; KANG *et al.*, 2018; CHEN *et al.*, 2018; WÜRFL *et al.*, 2018; ADLER; ÖKTEM, 2018; GONG *et al.*, 2018; ZHANG *et al.*, 2019; WU *et al.*, 2021). Machine-learning techniques that incorporate such priors are known as model-based deep learning (MBDL) (XIA *et al.*, 2023). Because MBDL accounts for specific domain knowledge, thanks to the availability of precise mathematical formulations proposed for specific tasks, the learning space is reduced and the CNNs can learn the data representation with fewer samples (SHLEZINGER *et al.*, 2023).

Inspired by an analytical model-based (MB) denoising pipeline proposed by our group (BORGES *et al.*, 2018), in this work we present an MBDL framework to restore low-dose DBT projections leveraging a signal-dependent Poisson-Gaussian noise model. The CNN is trained in a variance-stabilizing domain (VST) (STARCK; MURTAGH; BIJAOUI, 1998) where the noise is converted into approximately signal-independent, Gaussian-distributed, with unity variance. Furthermore, we explore a post-denoising image blending step which re-injects noisy data into the filtered one to recover part of the high-frequency components that are suppressed during denoising. Lastly, the approach of adding priors to the framework allowed the inclusion of a combined bias and residual variance loss function, which is used to adjust the trade-off between noise suppression and signal smearing in the restored image.

We hypothesize that the conversion of signal-dependent noise into signal-independent noise with fixed and known variance facilitates denoising as learning the local noise characteristics is no longer required. Furthermore, we conjecture that the image blending step may minimize excessive blurring and smearing from the noise suppression scheme. Lastly, we expect that having the ability to control the final noise characteristics through the combined bias-residual noise loss function may facilitate the potential clinical use as the network can be tuned to achieve a target image quality.

Considering the use of VST for denoising in a CNN context, Zhang *et al.* (ZHANG *et al.*, 2019) proposed a VST-Net to denoise natural images corrupted by Poisson noise. The network performs variance stabilization by adding a trainable layer to the CNN that converts signal-dependent Poisson noise into signal-independent noise. Compared to the work by Zhang *et al.*, our contributions include the use of an analytical form of the VST as opposed to a CNN-learned VST. Because we assume that the noise model is a prior, we can achieve reliable variance-stabilization using the Generalized Anscombe Transformation (GAT) (STARCK; MURTAGH; BIJAOUI, 1998), as well as reliable inverse transformation using the exact unbiased inverse transform (MAKITALO; FOI, 2012).

Moreover, the design of a bias-reducing form of the MSE using the concepts of bias and residual noise was explored by Narage *et al.* (NAGARE *et al.*, 2021). In their work, they use two independent samples of the same signal to back-propagate the bias-weighted MSE during the training of a denoising CNN. In our work, we use digital phantoms to perform fine-tuning of our CNN using the actual analytically estimated bias and residual noise. This is only possible because the fine-tuning step was performed using phantom images.

The remainder of this work is divided as follows. Section 5.3 gives a general overview of the image formation, the transformations, and a general overview of the proposed framework and its relation with the conventional mathematical formulations. Section 5.4 describes the network architectures, how the networks were trained, the loss functions used, the datasets, implementation details and finally the objective metrics. Section 5.5 presents and discusses the results and finally, Section 5.6 concludes the work, states the limitations of the work, and shows future works.

5.3 Theoretical background

Noise in DBT raw projections is typically described by a Poisson-Gaussian model (DOBBINS, 2000). The Poisson portion describes the quantum noise, which arises from the counting nature of X-ray generation and detection, whereas the Gaussian portion describes the electronic noise caused by thermal agitation (DOBBINS, 2000). Given a reference FD projection $\dot{z}$, each pixel of $\dot{z}$ can be formulated as

$$\dot{z}_i = \alpha_i p_i + n_i + \tau, \tag{5.1}$$

where i indicates the pixel spatial coordinate, $p_i \sim \mathcal{P}(\alpha_i^{-1} y_i)$ is a random variable with Poisson distribution, α_i is the spatially-dependent quantum gain, y_i is the (unknown) noise-free image, $n_i \sim \mathcal{N}(0, \sigma_e^2)$ is independent and identically distributed (i.i.d.) Gaussian noise, σ_e^2 is the variance of the electronic noise, and τ is the pixel offset. To facilitate the next mathematical steps, let us consider z as the linearized version of $\dot{z}$, which for this particular model translates into $z = \dot{z} - \tau$.

One can model an LD projection z^γ by fixing the radiographic factors as in z except for a reduction in the current-time product (mAs)

$$z_i^\gamma = \alpha_i p_i^\gamma + n_i, \tag{5.2}$$

where $p_i^\gamma \sim \mathcal{P}(\gamma \alpha_i^{-1} y_i)$ and $0 < \gamma < 1$ is the corresponding dose reduction factor.

Such dose reduction directly impacts the signal-to-noise ratio (SNR), such as, for z

$$SNR = \frac{E\{z_i \,|\, y_i\}^2}{var\{z_i \,|\, y_i\}} = \frac{y_i^2}{\alpha_i^{-1} y_i + \sigma_e^2}, \tag{5.3}$$

and for z^γ

$$SNR^\gamma = \frac{E\{z_i^\gamma \,|\, y_i\}^2}{var\{z_i^\gamma \,|\, y_i\}} = \frac{y_i^2}{\gamma^{-1} \alpha_i^{-1} y_i + \sigma_e^2 \gamma^{-2}}, \tag{5.4}$$

given that $0 < \gamma < 1$, thus $SNR^\gamma < SNR$, indicating that indeed the dose reduction causes a degradation in SNR.

In general, denoising frameworks are designed to achieve as low as possible residual noise variance, with the SNR being as high as possible. Thus, given a denoising framework $\mathfrak{D}$, ideally

$$var\{\mathfrak{D}\{z_i^\gamma\} \,|\, y_i\} \rightarrow 0 \tag{5.5}$$

$$SNR^{\mathfrak{D}} \rightarrow \infty \tag{5.6}$$

whereas in this work, inspired by the restoration framework proposed in Borges *et al.* (BORGES *et al.*, 2018), we aim to achieve

$$var\{\mathfrak{D}\{z_i^\gamma\} \,|\, y_i\} \rightarrow var\{z_i \,|\, y_i\} \tag{5.7}$$

$$SNR^{\mathfrak{D}} \rightarrow SNR \tag{5.8}$$

i.e., the image quality of our target denoised image should match the one of a standard FD acquisition instead of the quality of a noise-free signal y. The rationale for having the FD as the target is that aiming at a noise-free image might impose excessive smoothing and smearing to the underlying signal, which may negatively impact radiologists readings. Thus, having the FD as the target allows for a more conservative image denoising preserving relevant details.

The restoration process can be treated modularly, starting with a variance-stabilization module that converts the Poisson-Gaussian signal-dependent noise into approximately signal-independent Gaussian noise with unity variance. In the case of Poisson-Gaussian distributions, the GAT (STARCK; MURTAGH; BIJAOUI, 1998) is commonly used

$$f(z_i) = \begin{cases} 2\sqrt{\mathring{z}_i + \frac{3}{8} + \mathring{\sigma}_e^2}, & \text{if } \mathring{z}_i > -\frac{3}{8} - \mathring{\sigma}_e^2 \\ 0 & \text{if } \mathring{z}_i \leq -\frac{3}{8} - \mathring{\sigma}_e^2, \end{cases} \tag{5.9}$$

where

$$\mathring{z}_i = \frac{z}{\alpha_i} \qquad \mathring{\sigma}_e = \frac{\sigma_e}{\alpha_i}. \tag{5.10}$$

As the GAT is a non-linear transformation, noise removal may introduce bias if the inappropriate inverse transform is used (MAKITALO; FOI, 2012). Thus, the next denoising module consists of the closed-form approximation of the exact unbiased inverse (MAKITALO; FOI, 2012)

$$\frac{\widehat{y}_i^\gamma}{\alpha} = \frac{1}{4}\mathfrak{d}^2 + \frac{1}{4}\sqrt{\frac{3}{2}}\mathfrak{d}^{-1} - \frac{11}{8}\mathfrak{d}^{-2} + \frac{5}{8}\sqrt{\frac{3}{2}}\mathfrak{d}^{-3} - \frac{1}{8} - \mathring{\sigma}_e^2, \tag{5.11}$$

where $\mathfrak{d}$ is the denoised image in the GAT domain.

As our primary goal is to match the signal and noise properties of the restored and FD images, it is desirable that their expected mean and variance match. To that end, after denoising the image in the VST domain, the last denoising module consists of a weighted sum between the LD image z_i^γ and the denoised one $\widehat{y}_i^\gamma$ with weights defined by

$$\hat{z} = w_i z_i^\gamma + \bar{w}_i \widehat{y}_i^\gamma + \tau, \tag{5.12}$$

where

$$w_i = \sqrt{\frac{\alpha_i y_i + \sigma_e}{\gamma \alpha_i y_i + \sigma_e}}, \tag{5.13}$$

and

$$\bar{w}_i = \frac{1}{\alpha_i} - w_i, \tag{5.14}$$

are sufficient to make the restored image's expected mean and variance match the FD, considering the denoising was successful (BORGES *et al.*, 2018). As we do not have access to y, we approximate it to

$$\hat{y}_i = \frac{\widehat{y}_i^\gamma}{\gamma}. \tag{5.15}$$

In this work, the denoising modules defined by (5.9), (5.11) and (5.12) are implemented as layers of the CNN architecture. Note that even though they are implemented as layers, all their parameters are fixed, *i.e.*, they do not have any trainable weights.

Fig. 25 provides a general overview the proposed framework, where the GAT block implements (5.9), BS normalizes the signal, iGAT implements (5.11) and wAdd implements (5.12) and (5.13) with a residual layer.

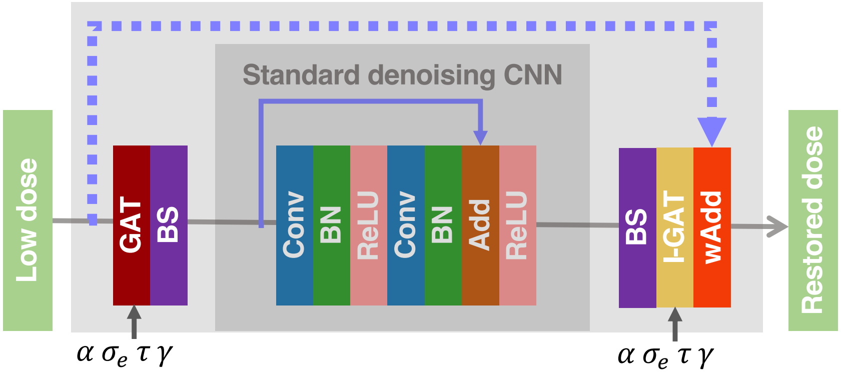


Figure 25 – Illustration of the proposed framework. Note that the framework can be applied independently of the network architecture and the illustrated standard denoising CNN is a generic architecture. GAT is the Generalized Anscombe Transformation layer, BS is a layer that normalizes the signal based on fixed values calculated from the entire dataset, Conv is a generic convolutional layer, BN is the commonly known batch normalization, ReLU is the activation function, Add is a generic residual layer, I-GAT the inverse of the GAT and wAdd is a layer similar to the residual layer but with calculated weights.

5.4 Materials & methods

5.4.1 Architecture

The denoising framework investigated in this work requires a “standard denoising CNN”, as shown in Fig. 25. Because our framework is agnostic to the denoising CNN, we tested three popular CNNs that are suitable for the denoising task: the Residual Network (ResNet) (HE *et al.*, 2016), the Residual Encoder-Decoder (RED) CNN (CHEN *et al.*, 2017a), and the U-Net (RONNEBERGER; FISCHER; BROX, 2015).

The ResNet contains 4 blocks and each block has 2 convolutional layers. The U-Net is composed of 3 encoder blocks and 3 decoders. Each encoder and decoder has 2 convolutional layers and the decoder has an additional convolutional transpose. The RED consists of 5 encoder blocks and 5 decoders with one convolutional and convolutional transpose, respectively for each block.

As a benchmark, we also performed the same denoising tasks using the three “standard denoising CNNs” without layers that implement (5.9), (5.11) and (5.12). Because those layers are not included in the benchmark, the standard denoising CNNs are not considered MBDL (model-based DL) pipelines, as they rely solely on the training data with no *a priori*. We address this class of CNNs as data-based deep learning (DBDL).

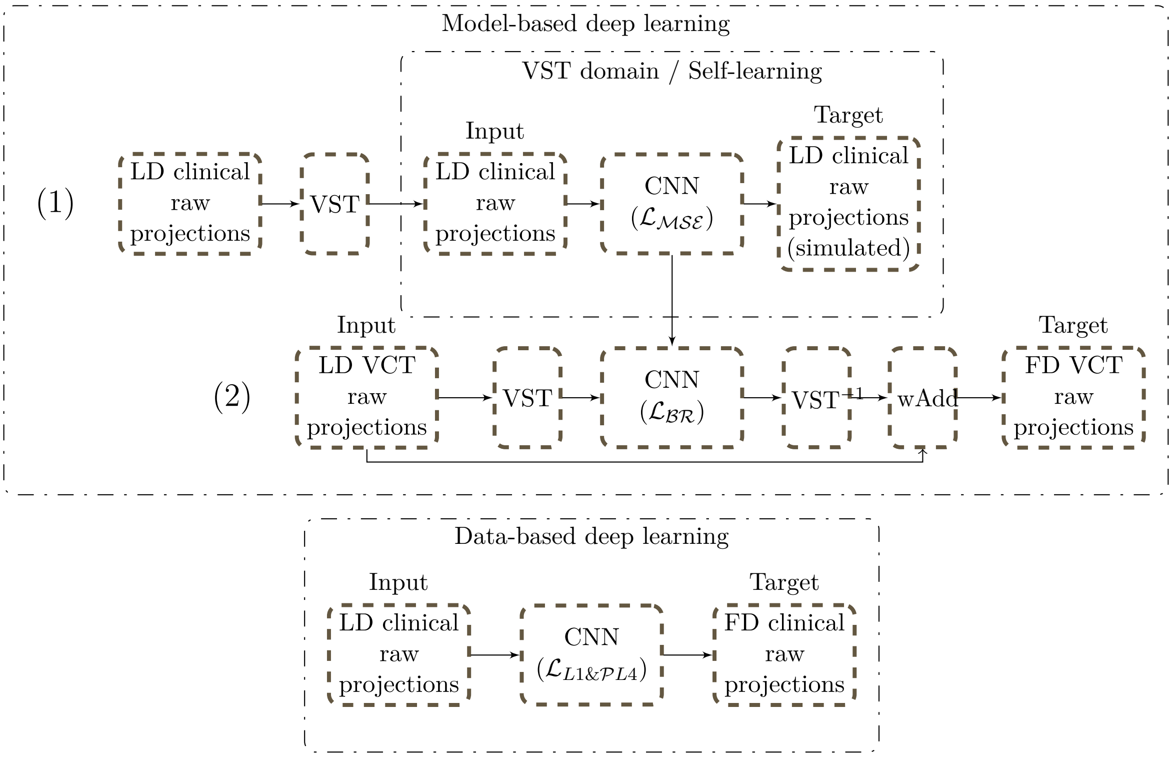

Model-based deep learning
VST domain / Self-learning
(1)
LD clinical raw projections
VST
Input
LD clinical raw projections
CNN ($\mathcal{L}_{\mathcal{MSE}}$)
Target
LD clinical raw projections (simulated)
(2)
Input
LD VCT raw projections
VST
CNN ($\mathcal{L}_{\mathcal{BR}}$)
VST^{-1}
wAdd
Target
FD VCT raw projections
Data-based deep learning
Input
LD clinical raw projections
CNN ($\mathcal{L}_{L1\&\mathcal{P}L4}$)
Target
FD clinical raw projections

Figure 26 – Scheme illustrating the conventional data-based deep learning framework for image restoration, where the model is meant to learn only from data. Also, the proposed model-based deep learning is separated into two steps. First (1), the model learns in a self-learning framework how to completely remove noise from an image in the VST domain. Here, the target images are simulated using the Noise2Sim framework. Second (2), the pre-trained network is used in the pipeline and can be fine-tuned using the proposed $\mathcal{L}_{\mathcal{BR}}$ loss function to adjust the denoising properties.

5.4.2 Training

For the training of the MBDL models, we adopted a two-stage strategy. In the first stage, the network was trained to estimate the noise-free signal in the VST domain using patient DBT data. Because we do not have access to noise-free patient images, we used a self-learning approach (LEHTINEN *et al.*, 2018), in particular, we chose the Noise2Sim (N2S) framework (NIU *et al.*, 2022). The N2S uses non-local image information to estimate a noise-free image in an unsupervised manner. Many other self-supervised methods are available and may be used as an alternative to N2S (HENDRIKSEN; PELT; BATENBURG, 2020; BATSON; ROYER, 2019; KRULL; BUCHHOLZ; JUG, 2019; MORAN *et al.*, 2020). It is important to emphasize that at this stage the CNN was trained to achieve a noise-free approximation of the stabilized signal, which is not the final goal of this paper, as defined in (5.12).

In the next stage, the pre-trained network is fine-tuned using a bias × residual-noise loss function. The final steps of inverse GAT (5.11) and image blending (5.12) are applied. Because bias and residual noise require several realizations of the same signal for reliable estimation, we used synthetic breast images for this training stage, which were simulated using a virtual clinical trial framework.

The benchmark DBDL methods were trained using raw DBT patient data. Because the final goal is to achieve FD images and not noise-free ones, we generated LD/FD pairs using a simulation framework (BORGES *et al.*, 2016; BORGES *et al.*, 2017). The training was done in two steps: in the first step, the $L1$ loss-function was used followed by fine-tuning using a perceptual-loss (JOHNSON; ALAHI; FEI-FEI, 2016) in the second step. This strategy showed superior performance if compared to other training strategies in a previous investigation from our group (SHAN *et al.*, 2023). In this scenario, self-supervised learning is not an appropriate approach as the CNN would be trained to achieve noise-free images.

Because the DBDL were trained with no extra layers, no *a priori* information was added to the denoising process, and thus the CNNs are expected to learn the restoration from the provided data only.

5.4.2.1 Loss functions

The importance of the loss function for natural image restoration was previously presented in (ZHAO *et al.*, 2016) and in (SHAN *et al.*, 2023) the authors explore the same subject in the field of medical imaging. In this work, we used different loss functions at different stages of the restoration, and below we list and define each of them and explain the rationale for its use.

- $\mathcal{L}_{\mathcal{MSE}}$: the mean of the squared error. We adopted this loss in the first stage of the training of our proposed framework, which aims at achieving a noise-free image.
- $\mathcal{L}_{\mathcal{BR}}$: this loss function explores the noise suppression versus blurring trade-off. The noise suppression is weighted by the residual noise variance while signal blurring and smearing are weighted by the signal bias. It is defined as

$$\mathcal{L}_{\mathcal{BR}} = \mathcal{B}^2 + \lambda_{\mathcal{R}} |\mathcal{R}_{FD} - \mathcal{R}_{RI}|, \tag{5.16}$$

  where $\lambda_{\mathcal{R}}$ is a weighting factor that controls the noise suppression versus blurring trade-off, $\mathcal{R}$ and $\mathcal{B}$ are the average local residual variance and average local bias estimated as described in (SHAN *et al.*, 2023). Higher $\lambda_{\mathcal{R}}$ will give higher importance to the residual noise matching between target and restored images, at the expense of a higher signal bias. Lower $\lambda_{\mathcal{R}}$ values will cause minimization of the signal bias while the mismatch between the target and restored residual noise might increase. This loss was used in the second stage of our proposed framework.

- $\mathcal{L}_{L1}$: the mean of the absolute error. This loss tends to minimize the noise at the expense of signal blurring (SHAN *et al.*, 2023). We used this loss in the first stage of the DBDL training.

- $\mathcal{L}_{PL4}$: the fourth layer of the perceptual loss CNN (JOHNSON; ALAHI; FEI-FEI, 2016). We used this loss in the second stage of the DBDL training aiming to retrieve image details that were lost during the first training stage.

5.4.3 Dataset

Because our methods require different datasets depending on the training strategies, we prepared three distinct datasets that are used at different stages of the experiments. Below we detail each of them.

5.4.3.1 Patient FD/LD pairs

The dataset consists of 1,982 retrospective exams (29,730 raw DBT projections) acquired at the Institute of Radiology (InRad), Faculty of Medicine, University of São Paulo (Brazil) under IRB approval (CAAE #56699016.7.0000.0065). The dataset contains patient cases acquired using the Hologic Selenia Dimensions equipment (Hologic, Bedford, MA). All images were carefully anonymized to preserve patients' medical records. The 1,982 exams were randomly selected from the hospital's PACS system to represent the screening population of that region.

We simulated the corresponding LD acquisitions using the method proposed at (BORGES *et al.*, 2016; BORGES *et al.*, 2017), which injects quantum and electronic noise in the VST domain. The LD images were simulated to mimic an acquisition with 50% of the exposure time, and thus 50% of the FD.

Because all the deep-leaning methods explored in this work are patch-based, we extracted 64 × 64 pixels ROIs from the full-size images. ROIs containing air were excluded, and 256,000 LD/FD ROI pairs were randomly selected from the remaining ones.

This dataset was used at two points of this work: the LD patches were used in the first training stage of our proposed pipeline, where the self-supervised approach was adopted. The FD/LD pairs were used in the training stages of the benchmark DL approaches (DBDL).

5.4.3.2 Synthetic FD/LD pairs

Synthetic DBT projections were generated using the OpenVCT virtual clinical trial framework, developed at the University of Pennsylvania (BARUFALDI *et al.*, 2018). Seven anatomies were simulated and the corresponding noise-free CC projections were created.

Note that, even though it is possible to simulate as many anatomies as desired, seven anatomies yielded a sufficient number of patches for this experiment.

We simulated five noise realizations of each projection set at FD and five realizations at LD using the method presented at (BORGES *et al.*, 2019). The FD radiographic factors were selected using the AEC table implemented in the OpenVCT pipeline, and the corresponding LD was defined as a 50% reduction in exposure time.

The geometry of the Hologic Selenia Dimensions system was used and the noise simulation accounts for the quantum and electronic noise as well as pixel cross-talk. A collection of 90,650 LD/FD ROI pairs, extracted from the VCT projections, was used in the second training stage of our proposed method so that we could generate several realizations of the same anatomy with different noise seeds and thus accurately estimate the residual variance and bias.

5.4.3.3 Phantom FD/LD pairs

As the goal of this work is to investigate noise suppression, the validation step should be performed using data with noise statistics that are as close as possible to what is found in a real clinical scenario. To that end, we acquired a dataset of anthropomorphic phantom DBT images using a Hologic Selenia Dimensions system installed at the Hospital of the University of Pennsylvania. The anthropomorphic phantom was designed at the University of Pennsylvania (CARTON *et al.*, 2011) and produced by CIRS, Inc. (Reston, VA).

We performed a total of 29 DBT acquisitions, split into 19 FD image sets and 10 image sets at 50% of the FD. The radiographic factors of the FD acquisitions were obtained using the automatic exposure control (AEC), which yielded 31 kVp and 60 mAs. We generated the corresponding LD projections by setting the system to manual mode and replicating the FD radiographic factors except for a reduction of 50% in the mAs.

This dataset was used as the testing sample for all methods investigated in this work, and thus all the results presented in the results section originated from this dataset. It is important to note that none of the data in this dataset was used by any of the methods during the training stages.

5.4.3.4 Uniform phantoms

Since we used several conventional mathematical equations as network layers, all necessary parameters such as $\mathring{\sigma}_e, \alpha, \tau$ and γ were estimated. We followed the same procedure and used the same uniform images as stated in (BORGES *et al.*, 2018). The same parameters were used for MBDL and MB restorations.

5.4.4 Experimental setup

All neural networks were implemented using the PyTorch Deep Learning library, trained and tested using an NVIDIA TITAN Xp with 12GB of RAM. Below we specify the experimental details for each of the frameworks.

Because our networks were trained using patches, the inference phase during the testing must also be performed using patches. To avoid border artifacts during image stitching, patches of size 192×192 were generated with overlap in all directions. After inference, the 64×64 central patch was used to reconstruct the full-resolution image.

5.4.4.1 MBDL

In the first training stage (self-supervised learning) the N2S was used in the VST domain as illustrated in Fig. 26, *i.e.*, the GAT shown in (5.9) was applied to the data before training. In this stage, we set the number of epochs to 70, with a batch size of 128, MSE loss function, 8 simulated images for the N2S and we maintained the default learning rate of the N2S, which starts at 0, goes up to 1×10^{-4} and returns to 0 following a cosine ramp. The dataset used to train the networks was the $256,000$ LD patches taken from patient cases, as described in 5.4.3.1.

In the second learning stage of the MBDL, we used the CNNs trained in the previous step and performed fine-tuning with the $\mathcal{L}_{\mathcal{BR}}$ loss function. The training was done for 5 epochs with a learning rate of 1×10^{-5}. At this stage, the MBDL model has the GAT, iGAT and wAdd layers incorporated into each architecture. Thus the loss function is now evaluated in the image domain and no longer calculated in the GAT domain as in the first training stage.

To estimate the $\mathcal{L}_{\mathcal{BR}}$ loss, we stacked the five noise realizations and set the batch size to 60. Note that we do not shuffle the batch after an epoch as we want to calculate the loss for patches with the same underlying signal. The loss was then calculated between the LD stack of patches and the corresponding FD stack, after the iGAT was applied, *i.e.*, in the projection domain. At this learning stage, as we need to estimate bias and residual noise variance, we adopted the synthetic FD/LD pairs dataset described in 5.4.3.2. We used the Adam optimizer with running averages of gradient and its square equal 0.5 and 0.999, respectively, in both training stages.

The proposed framework also relies on the appropriate selection of $\lambda_{\mathcal{R}}$, which controls the signal blurring versus noise suppression trade-off. To better understand the role of $\lambda_{\mathcal{R}}$ we performed a grid search investigation with the following setup:

- The three MBDL architectures were trained using 20 linearly spaced values of $\lambda_{\mathcal{R}}$ ranging from 0 to 2;

- The range of interest was detected (0 to 0.105) and a finer grid was created using 20 new linearly spaced $\lambda_{\mathcal{R}}$ values within the range of interest

5.4.4.2 DBDL

We trained the standard models as done in (SHAN *et al.*, 2023), *i.e.*, the first training stage used $\mathcal{L}_{L1}$ as loss function, with 60 epochs, batch size of 256 and learning rate starting at 1×10^{-4}, halved every 10 epochs.

In the second training stage the $\mathcal{L}_{\mathcal{P}L4}$ loss was used for 60 epochs, batch size of 64, the learning rate started at 1×10^{-5}, also halved every 10 epochs. We used the Adam optimizer with running averages of gradient and its square equal 0.5 and 0.999, respectively, in both training stages.

All DBDL networks were trained, in both stages, with $256,000$ LD/HD pairs of patches taken from the patient dataset described in 5.4.3.1.

5.4.4.3 MB

We also used the MB restoration technique presented in (BORGES *et al.*, 2018) as a benchmark. The algorithm relies on the accurate estimation of the noise parameters of a system to perform denoising in the VST domain. The denoiser used for this task was the block-matching and 3D filtering (BM3D) (DABOV *et al.*, 2007), and the GAT was the selected VST.

5.4.5 Figures of merit

In general, denoising methods involve a trade-off between noise suppression and signal blurring/smearing. One way to assess both aspects is through the estimation of bias and residual noise, by decomposing the MNSE, where bias quantifies the negative impact on the underlying signal while the residual noise measures the effectiveness of noise removal. When a noise-free image is the final goal of a denoising method, the residual noise is reduced as much as possible even if it involves extra penalties to the bias. On the other hand, in dose restoration, the aim is to achieve the same residual noise as the FD while keeping the bias as low as possible. In this work, we estimate bias ($\mathcal{B}$) and residual noise ($\mathcal{R}$) analytically, as done before in (BORGES *et al.*, 2018; SHAN *et al.*, 2023)

$$\mathcal{R} = \frac{1}{m \times n} \sum_{i}^{m \times n} \frac{\mathbb{V}\left(z_i'\right)}{y_i}, \tag{5.17}$$

$$\mathcal{B}^2 = \left[\frac{1}{m \times n} \sum_{i}^{m \times n} \frac{\left(\mathbb{E}\{z_i'\} - y_i\right)^2}{y_i}\right] - \frac{\mathcal{R}}{p}, \tag{5.18}$$

where m is the number of rows, n is the number of columns, $\mathbb{V}$ indicates the pixel-wise variance along the set of $z^{'}$ image realizations, $\mathbb{E}$ indicates the pixel-wise expectation, y is the noise-free image and p is the number of realizations in the $z^{'}$ set.

We evaluated the MNSE, as well as its decomposition into bias and residual noise, using the anthropomorphic phantom images described in 5.4.3.3. To generate the pseudo-ground-truth, we averaged 9 images acquired at FD. To estimate MNSE, ($\mathcal{B}$) and ($\mathcal{R}$) we used a separate, independent set of 10 realizations at FD and a set of 10 realizations at 50% of the dose. All metrics were evaluated inside the breast phantom region to avoid overestimation due to the uniform background. We adjusted the mean value of all images and restorations using an affine transformation between the input image and the pseudo-ground-truth to avoid errors due to small differences in DC level caused by aspects such as x-ray tube throughput and flat-fielding.

Since we are optimizing the MNSE decomposition in the proposed loss function, we also measured the structural similarity index (SSIM) (WANG *et al.*, 2004) and the SNR. We measured the SNR in all phantom images and its restorations as the ratio of the mean pixel value and its standard deviation along the five realizations. The metric was calculated only inside the breast region and an average filter of size 15×15 was used to smooth both the mean signal value and the standard deviation after their calculation. We used Scikit-image (WALT *et al.*, 2014) implementation of SSIM and calculated only inside the breast region. Note that we measured the SSIM of all images against the generated pseudo-ground-truth mentioned before.

For the sake of cross-validation, we measured all the metrics for each projection, *i.e.*, we first measured the metric for each projection and then calculate the statistics across the 15 projections. For both SNR and MNSE, we used different image realizations to calculate the metric, while in SSIM all realizations were used for statistics.

5.5 Results & discussions

In this section, we present the visual comparisons and the quantitative results for both MBDL and DBDL approaches. We show the results as an ablation study, *i.e.*, first, the DBDL models are presented such that *no priors* are added and the network only learns from data. We also present intermediate results of the MBDL before the fine-tuning, *i.e.*, the VST is applied, introducing the *a priori* on the noise model, however, the bias versus residual noise fine-tuning was not done at this stage. We refer to the intermediate results as MBDL. Finally, we present the MBDL$^{\lambda}$ results with both the VST and the λ tuning. For the sake of simplicity, we only show results related to the selected point for λ.

5.5.1 $\lambda_{\mathcal{R}}$ Optimization

As previously mentioned, the impact of $\lambda_{\mathcal{R}}$ was investigated using a grid search approach. Fig. 27 shows the results of $\mathcal{B}$ and $\mathcal{R}$ for different $\lambda_{\mathcal{R}}$ values. In these plots we can observe the trade-off between bias and residual noise, *i.e.*, as the $\lambda_{\mathcal{R}}$ increases, more importance is given to the matching of $\mathcal{R}$, causing an extra penalty to the bias. On the other hand, at low $\lambda_{\mathcal{R}}$ the network prioritizes the underlying signal, resulting in lower bias and higher noise variance. In DBT image restoration, we seek a match between the $\mathcal{R}$ of FD and restored images, with bias as low as possible. The choice of $\lambda_{\mathcal{R}}$ in which the CNN should operate is quite dependent on the task at hand. In this work, we chose the $\lambda_{\mathcal{R}}$that yields a $\mathcal{R}$ within a 6% margin of the FD $\mathcal{R}$. This condition was met at $\lambda_{\mathcal{R}} = 0.025$ for the RED architecture, at $\lambda_{\mathcal{R}} = 0.421$ for the ResNet architecture and at $\lambda_{\mathcal{R}} = 0.035$ for the U-Net architecture.

5.5.2 Visual comparison

Fig. 28 illustrates the restoration of LD projections by different approaches. The first row shows an ROI of the input image, the MB restoration and the target FD, respectively. The following rows show the DL methods for each architecture. Column (a) illustrates the restoration through the DBDL approach, *i.e.*, using both $\mathcal{L}_{L1}$ and $\mathcal{L}_{\mathcal{P}L4}$ losses. Column (b) shows the proposed MBDL using self-supervised learning and column (c) MBDL$^{\lambda}$ for the selected point discussed in Section 5.5.1. Note, this ROI was extracted from the central projection of the anthropomorphic breast phantom and they are plotted in the same window and level.

In terms of DBDL restoration, $\mathcal{L}_{L1}$ loss is first used and removes almost all the noise, smoothing the underlying signal as well. This is not desirable since important features can be removed in the process. The $\mathcal{L}_{\mathcal{P}L4}$ is then used trying to retrieve the image details. It is possible to note in Fig. 28 (a) that the network can reduce the noise while keeping the high-frequency components of the underlying signal. However, some artifacts can be seen in the results. In special, RED architecture is visually less noisy than the others.

For the proposed framework, *i.e.*, MBDL, it is possible to notice the network trained with the self-supervised learning approach (b) with the GAT layers was able to successfully reduce the noise from the input and keep the image details. N2S by itself benefited from training in the VST domain, where noise has approximately unit variance and already achieved good results. Although we used this framework for training the network, it is possible to use any other architecture and framework for denoising. The limitation of this method is that it is not possible to control the network operation point in terms of bias and $\mathcal{R}$. It is even noticeable in the ROIs that all architectures have higher residual noise compared to the MB and the FD ones.

In MBDL$^{\lambda}$, since we selected a point that is within 5% of the FD residual noise, it

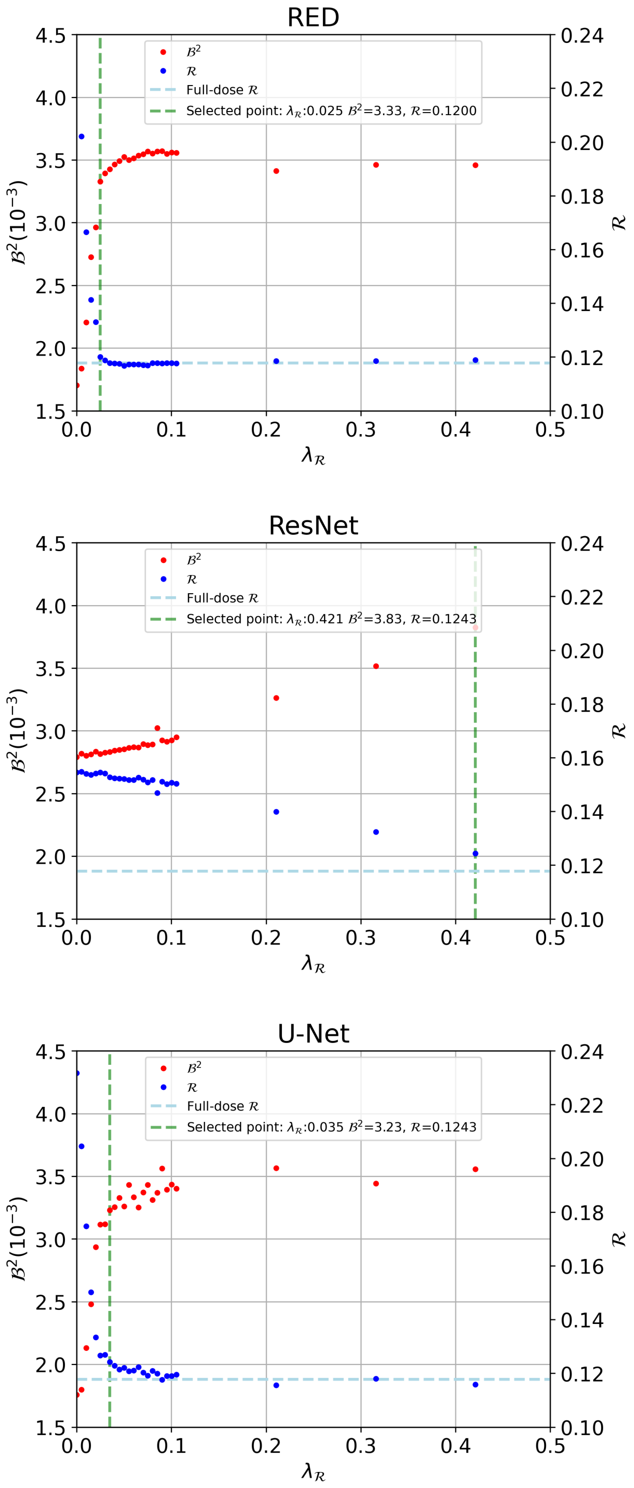


Figure 27 – $\lambda_{\mathcal{R}}$ tuning plot on the last epoch. The graphic illustrates the network operation points in terms of bias and $\mathcal{R}$ for each $\lambda_{\mathcal{R}}$ value.

is possible to see that for all architectures the noise visual aspect is very close to the FD, at the expense of slightly increasing the artifacts. All architectures were able to successfully reduce the noise from the input since their starting point was the self-supervised model. This point makes the network compromised in terms of preservation of the underlying signal and $\mathcal{R}$ matching, *i.e.*, the loss is balanced between error and noise variance matching.

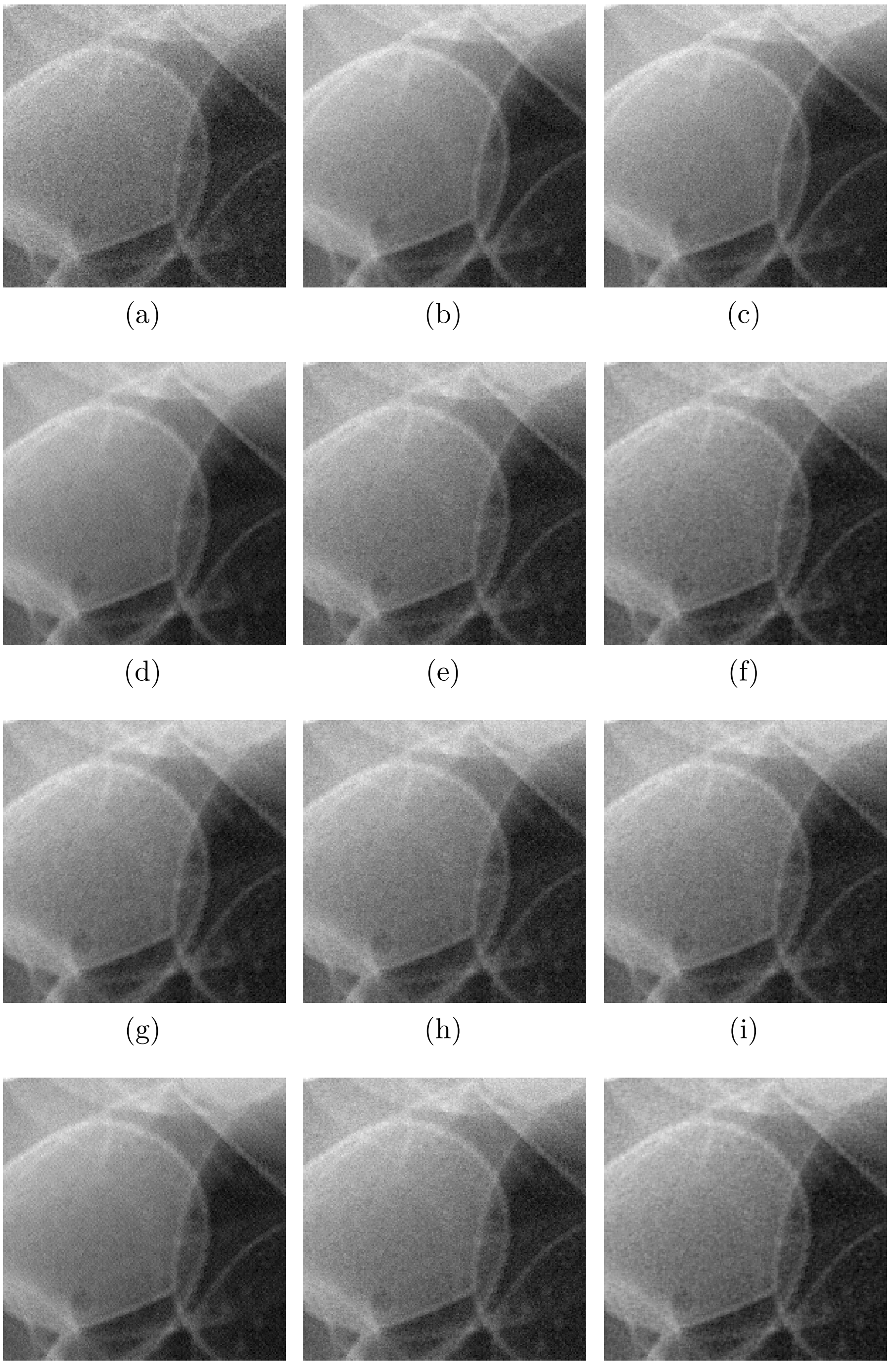


Figure 28 – Visualization of an ROI for a visual comparison of the DBDL and MBDL restorations for different architectures. Where (a) represents the LD image, (b) the corresponding FD image and (c) the MB method restoration. Figures (d) to (f) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the RED architecture. Figures (g) to (i) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the ResNet architecture. Figures (j) to (l) illustrate DBDL, MBDL and MBDL$^{\lambda}$, respectively, for the U-Net architecture. Note that each line represents a different architecture.

We also plotted the MB restoration as a benchmark for all the proposed restoration methods.

5.5.3 Quantitative analysis

Table 15 provides an objective analysis, in terms of MNSE and its decomposition on bias and residual noise, for LD and FD acquisitions, DL restorations and also the MB method. We highlighted, within the DL methods, the lowest value of bias and the closest $\mathcal{R}$ to the FD value. Again, our goal is to compare different training strategies and frameworks and not different CNN architectures.

Table 15 – MNSE analysis with its decomposition on bias and $\mathcal{R}$ for the DBDL and MBDL approaches and comparison with LD, FD and MB results. Values highlighted mean the lowest bias for each architecture. We also highlighted the closest $\mathcal{R}$ to the value from FD images.

		MNSE	$\mathcal{R}$	$\mathcal{B}^2(10^{-3})$
LD		0.250 ± 0.002	0.2485 ± 0.0010	1.7 ± 1.8
FD		0.118 ± 0.001	0.1178 ± 0.0006	0.7 ± 0.5
MB		0.135 ± 0.002	0.1322 ± 0.0006	2.9 ± 1.8
DBDL	RED	0.108 ± 0.002	0.1039 ± 0.0003	4.4 ± 1.9
	U-Net	0.120 ± 0.002	0.1123 ± 0.0003	7.2 ± 1.9
	ResNet	0.120 ± 0.002	0.1111 ± 0.0004	8.5 ± 1.9
MBDL	RED	0.144 ± 0.002	0.1414 ± 0.0006	**2.9** ± 1.8
	U-Net	0.145 ± 0.002	0.1414 ± 0.0006	3.3 ± 1.8
	ResNet	0.145 ± 0.002	0.1419 ± 0.0005	3.5 ± 1.8
MBDL$^{\lambda}$	RED	0.123 ± 0.001	**0.1200** ± 0.0004	3.3 ± 1.6
	U-Net	0.128 ± 0.002	0.1243 ± 0.0007	3.2 ± 1.9
	ResNet	0.128 ± 0.002	0.1243 ± 0.0005	3.8 ± 2.0

Looking only at the MNSE, one could say that the RED architecture in DBDL mode achieved the best result. This is not what we see in the ROIs above, since the results were over-smoothed for this specific architecture. Looking at the bias and the residual noise, in general, MB performed better compared with the DL method in the DBDL framework. We also can see that in this mode, there is no control over how the network behaves regarding denoising properties since the $\mathcal{B}$ and $\mathcal{R}$ have great variability within different architectures.

In the MBDL framework, we can notice that all architectures achieved the same residual noise. This demonstrated the stability of the training process in the VST domain even with different architectures. However, compared to the FD, those values are within a 20% margin of the residual noise. When we look at the benchmark, its residual noise is within a 12.2% margin. Even though residual noise was virtually the same, different architectures yielded different biases, illustrating the different characteristics of distinct architectures.

For MBDL$^{\lambda}$, as a consequence of the selected point, all residuals are within the 6% margin. The closest value, highlighted in bold, is present in this framework. We can notice a slight increase in bias for both RED and ResNet. For the U-Net, the bias was kept virtually the same. These facts demonstrate the effectiveness of the $\mathcal{L}_{\mathcal{BR}}$ loss function to fine-tune the network. Also, as shown in Fig. 27, with this loss it is possible to control the network operation concerning denoising strength, accomplishing the specific results, unlike the DBDL mode. Moreover, adding prior knowledge to the training process enhanced the restoration quantitative results, as shown by MBDL with self-supervised learning, using the N2S framework.

One can notice that the bias standard deviation is high compared to the mean value, however, this high value is also present in the LD acquisition, suggesting a high variance in the bias within different DBT projections.

Fig. 29 illustrates the $\mathcal{B}$ and $\mathcal{R}$ relation for all restoration methods. The dashed gray lines indicate points where the MNSE has the same value. Even though the DBDL method with ResNet and U-Net have approximately the same MNSE to the FD image, *i.e.*, they are almost in the same line, these restorations are not desirable in the clinical routine since they have high values of bias. It is worth noting that MBDL with self-learning (lozenge) all live very close to each other, confirming the stability of the framework across different architectures, and they are also close to the MB restoration, demonstrating the effectiveness of the method. When looking at the MBDL$^{\lambda}$, they were all capable of decreasing the residual noise, without the expense of highly increasing the bias term. Ideally, we want a point to be as close as possible to the FD one. However, is still not clear the relation between residual noise and bias.

Since were are optimizing the MNSE decomposition, it is worth exploring different objective metrics. Table 16 shows the assessment of SSIM and SNR for all different methods. We can notice that DBDL has higher SSIM and SNR values compared to the FD. In terms of SNR, we can argue that they are performing more denoising, decreasing the standard deviation and consequently increasing the SNR. For the MBDL, both SSIM and SNR are smaller than the FD one. For the SNR, since their residual noise is higher, it is expected to have a smaller SNR due to the high standard deviation. Again, it is worth noting the similarity of the metrics within different architectures, reinforcing the framework stability. When it comes to MBDL$^{\lambda}$, this framework was able to increase both SSIM and SNR, compared to the MBDL without fine-tuning. However, with levels close to the FD reference and also to the MB one.

Another advantage of MBDL mode is the training time, as illustrated by Table 17. For DBDL, the fastest model, *i.e.*, the ResNet, more than 86h were spent for training purposes, among 7.5h for pre-training with $\mathcal{L}_{L1}$ and 79 for training with $\mathcal{L}_{\mathcal{P}L4}$. MBDL only takes approximately 9 hours for training in the self-supervised framework, as it uses the

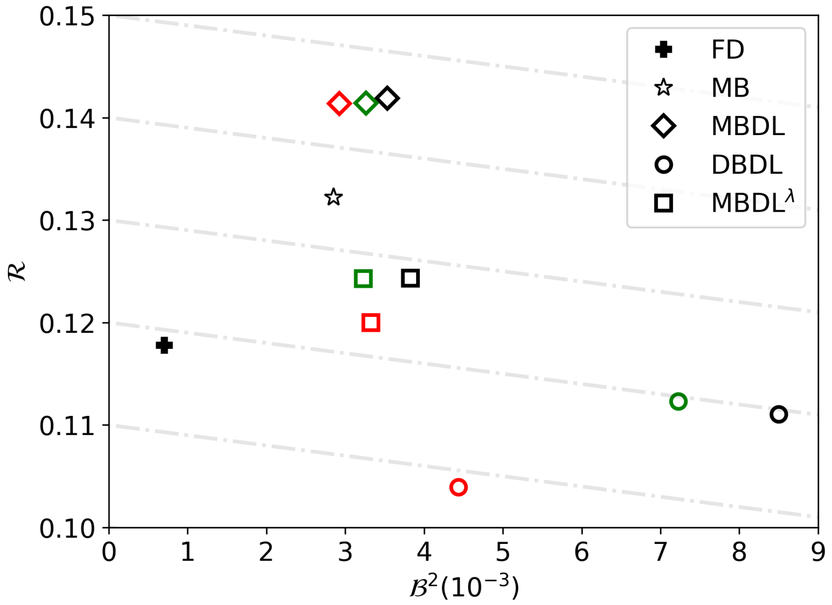


Figure 29 – Illustration of bias and residual noise for each restoration method. The gray dashed lines show points where the MNSE is the same. The green is related to the U-Net architecture, red with the RED and black with the ResNet.

Table 16 – SSIM and SNR objective metrics for the DBDL and MBDL approaches and comparison with LD, FD and MB results. We highlighted the closest SSIM and SNR to the FD value.

		SSIM	SNR
LD		0.9684 ± 0.0004	55.4 ± 0.4
FD		0.9839 ± 0.0002	78.7 ± 0.5
MB		0.9824 ± 0.0002	75.6 ± 0.5
DBDL	RED	0.9851 ± 0.0002	85.2 ± 0.5
	U-Net	0.9837 ± 0.0002	82.0 ± 0.5
	ResNet	0.9848 ± 0.0002	82.3 ± 0.5
MBDL	RED	0.9812 ± 0.0002	73.2 ± 0.5
	U-Net	0.9812 ± 0.0002	73.2 ± 0.5
	ResNet	0.9811 ± 0.0002	73.0 ± 0.5
MBDL$^\lambda$	RED	0.9845 ± 0.0002	**79.2** ± 0.5
	U-Net	**0.9838** ± 0.0002	77.9 ± 0.5
	ResNet	0.9836 ± 0.0002	77.8 ± 0.5

low-time cost MSE loss function. Finally, the fine-tuning process with $\mathcal{L}_{\mathcal{BR}}$ loss only takes at the most half an hour, totalizing 10h at the most. When it comes to inference time, both DBDL and MBDL are generally the same, since only the GAT, its inverse and the weighted sum are added to the framework It is also worth noting that $\mathcal{L}_{\mathcal{P}L4}$ requires the VGG network to be loaded in GPU memory, demanding more memory and, consequently decreasing image batch size.

Table 17 – Training time, in hours, for DBDL, MBDL and MBDL$^{\lambda}$ modes. Inference time for all frameworks, in seconds

Framework	Architecture	Training (h)	Inference (s)
DBDL	RED	92	5.21
	U-Net	115	4.99
	ResNet	86.5	2.22
MBDL	RED	9.6	5.23
	U-Net	9.1	5.08
	ResNet	8.5	2.22
MBDL$^{\lambda}$	RED	9.9	5.23
	U-Net	9.7	5.08
	ResNet	8.7	2.22

5.6 Conclusion

In conclusion, CNNs that incorporate prior knowledge on its architecture derive advantages from it, yielding superior results compared to those that lack such knowledge. When MBDL was evaluated against DBDL mode, it achieved superior results in training time, bias, and residual noise and also required less GPU memory. Also, all objective metrics were comparable to our benchmark, demonstrating the effectiveness of the proposed framework.

This work has some limitations. We restricted the network architecture to common models. We reinforce that the proposed framework does not depend on the architecture itself so the better the architecture the better the results are. Further evaluation with state-of-the-art architectures such as Transformers is valid. Also, different self-learning methods can be used for training the network in the VST domain. Another limitation respects the equipment parameters estimation necessity for the GAT and weighted sum layers. Future works could investigate performing the GAT, its inverse and the wSum as trainable layers and compare the results. Also, there is a necessity to inspect the $\lambda_{\mathcal{R}}$ parameter in terms of lesion diagnosis with radiologists. This would answer the question of whether a low $\mathcal{B}$ image with a slightly high $\mathcal{R}$ is preferred or an $\mathcal{R}$ match between the restored image and FD, with a loss of $\mathcal{B}$.

Finally, we claim that the presented framework for training networks for image restoration can be expanded for other areas in the medical image field and also for different areas such as natural images and video processing.

Acknowledgment

This work was supported in part by the São Paulo Research Foundation (FAPESP grant 2021/12673-6), National Council for Scientific and Technological Development (CNPq), the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001), and National Institute of Health (R01EB026646, R01CA233888, R01CA237267, and R01HL151561).

The authors would like to thank the Institute of Radiology (InRad), in particular, Dra. Denise Yanikian Nersissian and Éric Francisco Scolastici, for providing the clinical data for training. The authors also would like to thank Dr. Andrew D. A. Maidment, from the University of Pennsylvania, for making the anthropomorphic breast phantom images available for this work.

6 PAPER 5:SUPPRESSING NOISE CORRELATION IN DIGITAL BREAST TOMOSYNTHESIS USING CONVOLUTIONAL NEURAL NETWORK AND VIRTUAL CLINICAL TRIALS

The material presented in this chapter was published in: Vimieiro, Rodrigo de Barros, Lucas Rodrigues Borges, Renato França Caron, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Ge Wang, and Marcelo Andrade da Costa Vieira. “Suppressing noise correlation in digital breast tomosynthesis using convolutional neural network and virtual clinical trials.” In 16th International Workshop on Breast Imaging (IWBI2022), vol. 12286, p. 314-320. SPIE, 2022. Available at: https://doi.org/10.1117/12.2625357. The author obtained permission for total and/or partial reproduction of the content from the publisher SPIE.

6.1 Abstract

It is well-known that x-ray systems featuring indirect detectors are affected by noise spatial correlation. In the case of digital breast tomosynthesis (DBT), this phenomenon might affect the perception of small details in the image, such as microcalcifications. In this work, we propose the use of a deep convolutional neural network (CNN) to restore DBT projections degraded with correlated noise using the framework of a cycle generative adversarial network (cycle-GAN). To generate pairs of images for the training procedure, we used a virtual clinical trial (VCT) system. Two approaches were evaluated: in the first one, the network was trained to perform noise decorrelation by changing the frequency-dependency of the noise in the input image, but keeping the other characteristics. In the second approach, the network was trained to perform denoising and decorrelation, with the objective of generating an image with frequency-independent (white) noise and with characteristics equivalent to an acquisition with a radiation exposure four times greater than the input image. We tested the network with virtual and clinical images and we found that in both training approaches the model successfully corrected the power spectrum of the input images. Our code is available at www.github.com/LAVI-USP/NoiseDecorrelation-DBT.

6.2 Introduction

In general, digital breast tomosynthesis (DBT) systems use a large flat panel detector either using the technology of amorphous silicon (a-Si) coupled with a thallium-doped cesium iodide (CsI:Tl) or amorphous selenium (a-Se) (GHETTI *et al.*, 2008; SALVAGNINI *et al.*, 2013). The former contains layers of phosphor, a-Si and a thin film transistor (TFT), which convert x-rays to visible light and then to an electrical signal. The latter consists of an a-Se and a TFT layer (MCENTEE, 2017).

The indirect detection system is commonly known to have considerate noise spatial correlation due to the scintillation of the phosphor layer. It was reported on a recent work of Boita *et al.* (BOITA *et al.*, 2021) that noise correlation has a direct impact on the detection of subtle structures in digital mammography. In worse cases, noise correlation can create an undesired signal which could resemble tinny microcalcifications. Also, many restorations methods consider that the input image was corrupted by frequency-independent (white) noise. The fact that indirect detectors impose noise correlation might affect the performance of such methods.

In this work, we propose a deep convolutional neural network (CNN) to restore DBT projections that were degraded by frequency-dependent (correlated) noise. We used a cycle generative adversarial network (cycle-GAN) (ZHU *et al.*, 2017) in the restoration process. This network was initially used to transform images of horses into images of zebras and vice-versa. In this work, two medical imaging restoration approaches were considered. In the fist approach, the network was trained to perform noise decorrelation, *i.e.*, to transform the correlated noise in the input image into white noise in the output image. In the second approach, the network was trained to perform noise decorrelation and also to suppress the noise to a level equivalent to an acquisition with a radiation exposure four times greater than the input image.

6.3 Materials & methods

6.3.1 Training and testing datasets

Even though it is possible to acquire a clinical image dataset to train deep neural networks, as we have done previously in Shan *et al.* (SHAN *et al.*, 2023) and Vimieiro *et al.* (VIMIEIRO *et al.*, 2022), it is not possible to obtain labels without noise correlation, especially for systems with indirect detectors. Thus, for our training dataset, we simulated DBT projections using the virtual clinical trial system (OpenVCT), developed by the University of Pennsylvania (BAKIC *et al.*, 2017).

6.3.1.1 Training

Five hundred 3D breast phantoms with different anatomical structures were simulated using OpenVCT. Then, we generated three sets of raw projections for all the volumes. The first dataset, named independent full-dose (IFD), was generated considering the radiographic factors of a standard-dose acquisition with 90mAs and corrupted with no-correlated (white) noise. The second dataset, named correlated full-dose (CFD), was generated considering the same radiographic factors of the previous one, but the images were corrupted with correlated noise. The noise correlation kernel was estimated from a DBT system with a-Si(CsI:Tl) detector. Third, a four-fold acquisition with no-correlated noise was considered, *i.e.*, the same radiographic factors of the previous ones but with

4x longer exposure time (360mAs). This dataset was named independent four-fold (IFF). Fig. 30 illustrates how the projection data was organized for the training scheme. We followed the work of Borges *et al.* (BORGES *et al.*, 2019) for the noise simulation.

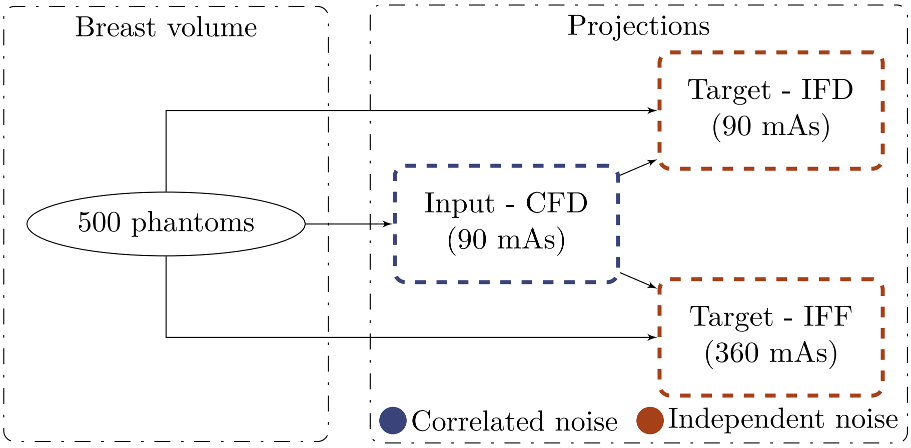


Figure 30 – Scheme illustrating how the training dataset was generated. From the 500 phantoms, three sets of projections were generated. The input dataset at 90mAs with independent white noise and two targets with correlated noise. One target at the same dose as the input and the other one at 360mAs.

The datasets CFD and IFD were used to train the first proposed network (cGAN #1), where the approach is to perform only noise decorrelation, without changing the total amount of noise of the image. The datasets CFD and IFF were used to train the second network (cGAN #2), where the approach is to perform noise decorrelation and denoising.

The training dataset contained 256,000 patches of size 64 × 64 pixels, extracted randomly from the raw DBT projection images. The network input was composed of 90mAs correlated-noise patches and the labels were either independent-noise patches at 90mAs or 360mAs.

6.3.1.2 Testing

For testing, we selected a virtual phantom that was not used in the training step and generated 10 acquisitions of it. For objective validation, we measured the mean normalized squared error (MNSE) decomposed into residual noise ($\mathcal{R}_{\mathcal{N}}$) and bias squared ($\mathcal{B}^2$). This decomposition allows the evaluation of the restoration methods in terms of signal smoothing and residual noise separately. This metric was proposed by Borges *et al.* (BORGES *et al.*, 2018) and further information is available in Shan *et al.* (SHAN *et al.*, 2023). All 10 acquisitions were utilized to measure the MNSE comparing with the ground truth (GT) from the OpenVCT software. We also measured the power spectrum (PS) as an objective metric (KAVURI; DAS, 2020), as follow:

$$PS = \frac{p_s^2}{n_r\, n_p} \sum_{k}^{n_r} \left|\mathcal{FFT}\{h \times u_k\}\right|^2 , \tag{6.1}$$

$$NPS = \frac{PS}{LAS^2}, \tag{6.2}$$

where n_r is the number of extracted region of interests (ROI), n_p is the ROI total number of pixels, p_s is the pixel size in mm, h is a Hanning window, u_k is the extracted ROI and $\mathcal{FFT}$ indicates the discrete fast Fourier transform. We normalized the PS through (6.2) by the LAS, which is the large are signal, calculated as the squared mean pixel value inside the segmented breast. The 1D plot was calculated as a radially mean of the 2D spectrum.

Finally, we tested the network on real clinical images. We selected six images from a dataset we collected in cooperation with the Barretos Cancer Hospital (Brazil). These images are from 6 different patients acquired in the Senographe Pristina (GE) DBT system under IRB approval *CAAE #986670.* This DBT system contains an a-Si(CsI:Tl) detector. All these images were acquired using the automatic exposure control (AEC) from the system and were fully anonymized to preserve patients' medical records.

6.3.2 Network and implementation details

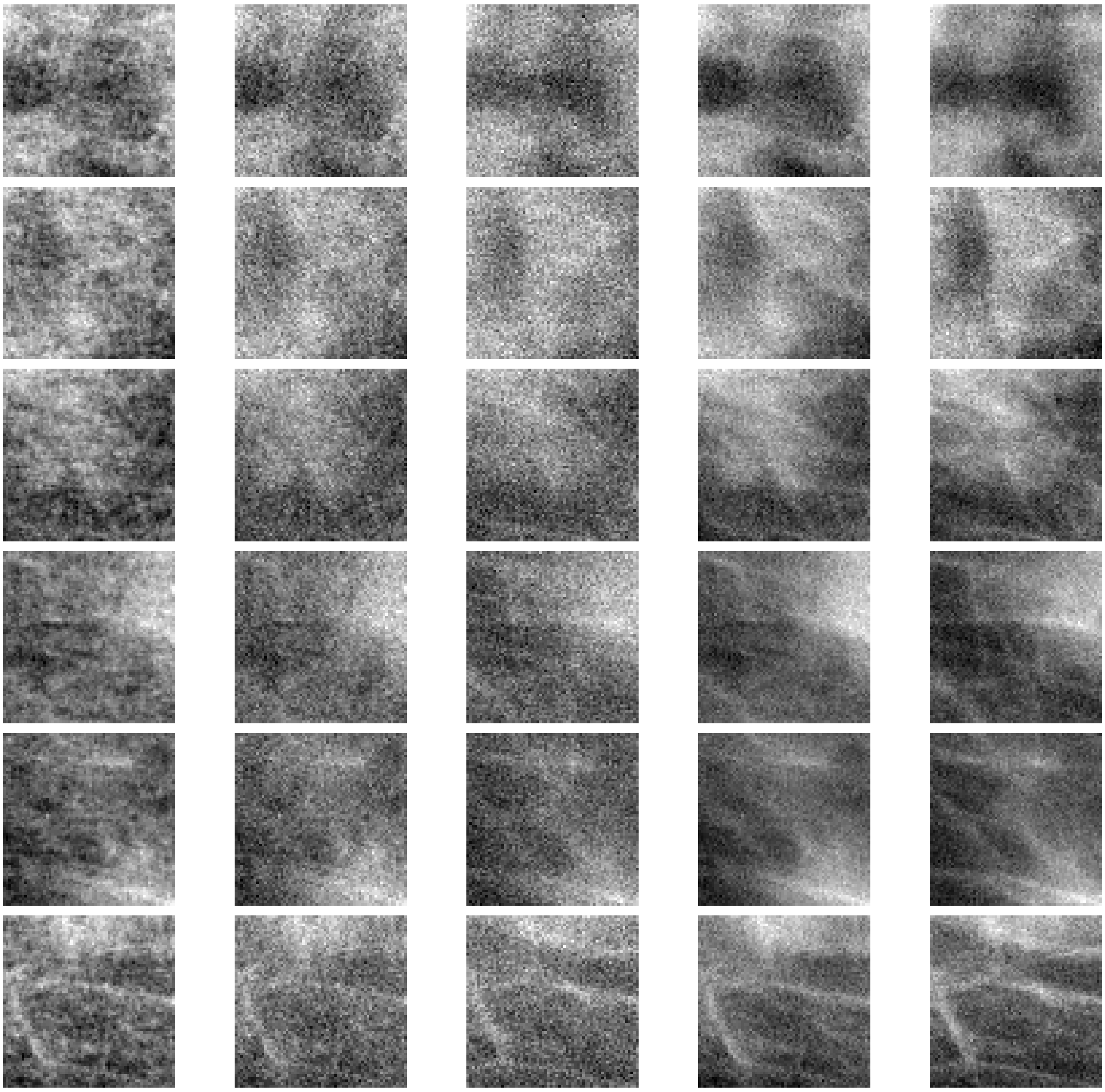
For the network architectures, we used the model proposed by Shan *et al.* (SHAN *et al.*, 2023) in both generators. For the discriminators, we kept the same architecture of the original cycle-GAN proposed by Zhu *et al.* (ZHU *et al.*, 2017). Also, we used the same losses and parameters stated in the original work. The CNN was implemented using the PyTorch Deep Learning library[1], trained and tested using an NVIDIA TITAN Xp with 12GB of RAM. All networks were trained for 30 epochs.

6.4 Results & discussions

Fig. 31 illustrates several ROIs extracted from the virtual phantom used for testing. The first column (a) illustrates the ROIs which were input for both proposed networks. Respectively, columns (b) and (d) are the network outputs for the cGAN #1 and #2, respectively. Columns (c) and (e) are the target for the cGAN #1 and #2, respectively. Here again, we can see that cGAN #1 could achieve noise results that resemble an independent noise, however, the resulting image looks noisier than the input. For the cGAN #2, the results are very close to the target one. The fine details are more sharp compared to the input. In fact, some structures do not exactly match the target one since those details were lost in the input image due to the system degradation simulation. Although these results are from synthetic images, we claim that the networks did not overfit to this type of data, as we demonstrated on the real clinical results from Fig. 32.

Fig. 32 illustrates ROIs of the central projection for the DBT clinical images. The first row represents the original images, the second row the results of the network targeting an independent noise at the same radiation dose (cGAN #1) and the third row the results

[1] www.pytorch.org



(a) Input-90mAs (b) cGan#1-90mAs (c) Target#1-90mAs (d) cGan#2-360mAs (e) Target#2-360mAs

Figure 31 – ROIs of the testing phantom generated with the OpenVCT. Each line is showing a region of interest extracted from: (a) input images at the FD with correlated noise; (b) cGAN #1 restoration at the FD; (c) target FD images with independent noise; (d) cGAN #2 restoration at 4x the FD and (e) target images at 4x the FD with independent noise.

of the network targeting a four-fold dose with independent noise (cGAN #2). In the second row, we can see that the network was able to decorrelate the noise but at a cost of increasing the overall noise in the images. For the third row, the image contains less relative noise and also with a white noise appearance. The results are more evident on the columns (b) and (e), where the structures look even sharper than the original input. Still on the third row, although some microcalcifications were blurred, the network did not remove any of them. As we do not have the GT for these clinical images, we might not be able to distinguish whether some small white spots are real microcalcifications or the result of noise correlation, especially on the (f) column.

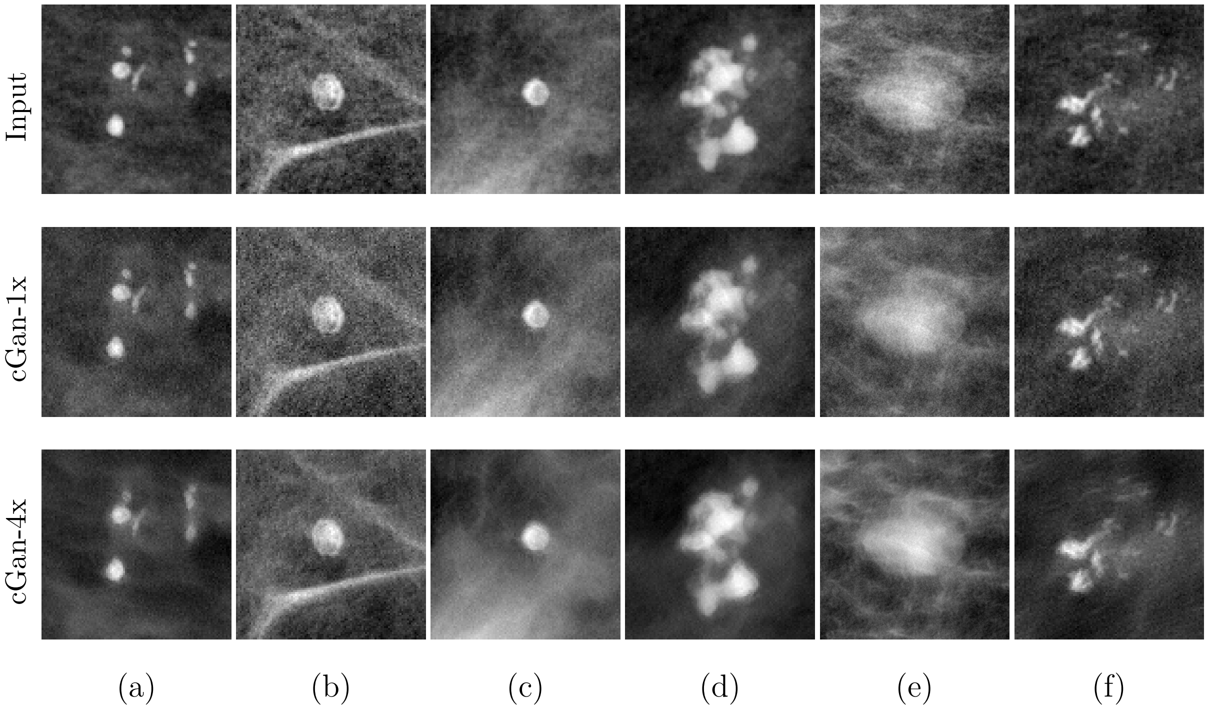
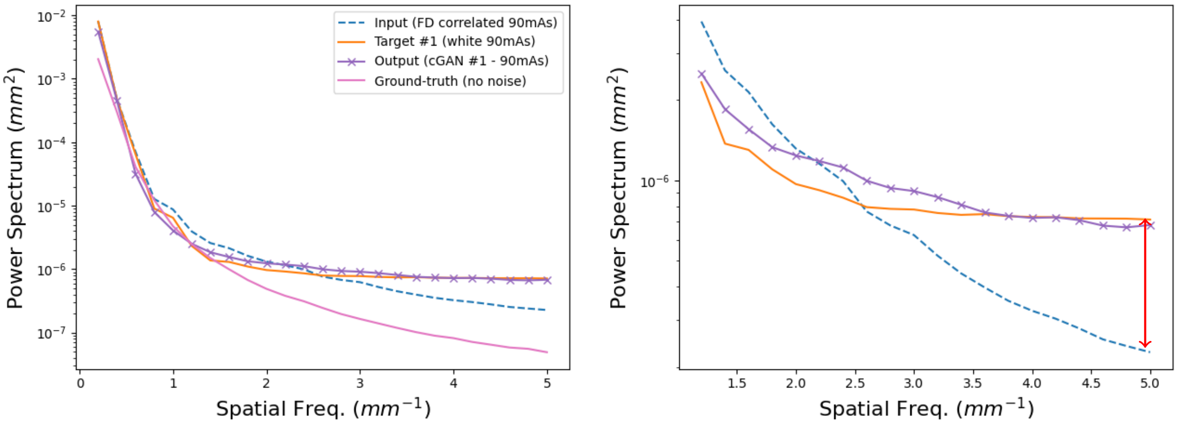


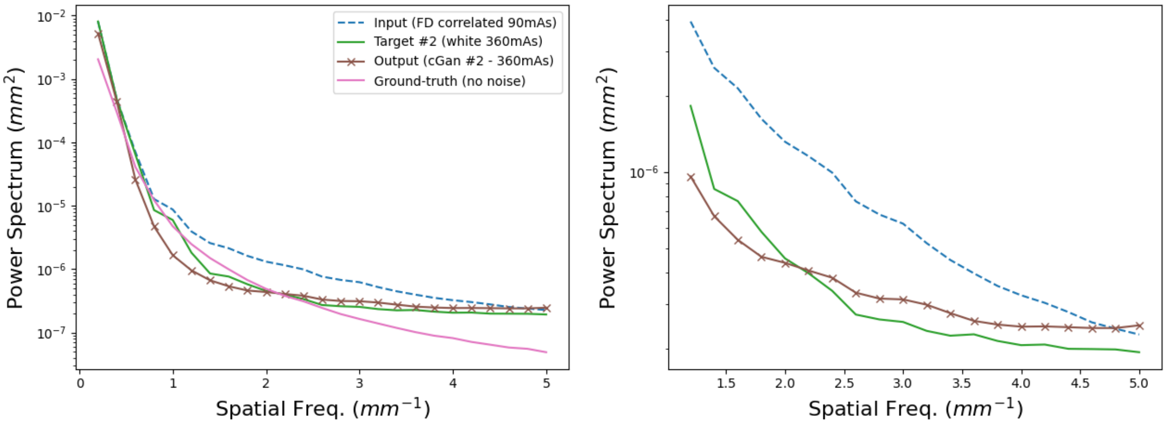
Figure 32 – ROIs extracted from DBT central projections of six clinical cases. The first row represents the input images at the standard full dose. The second and third rows represent the output from the networks targeting the independent noise at the same dose (cGAN #1) and four times the dose (cGAN #2), respectively. (a) to (f) illustrate the six different clinical cases.

Fig. 33 illustrates the PS measurements in the testing phantom for the cGAN #1. The blue dotted line shows the curve for the standard FD image with correlated noise, *i.e.*, the network input. In this task, as described before, the network has as target the PS illustrated in the orange curve, *i.e.*, the network was responsible to perform only the decorrelation task. It is known that anatomical noise prevails in the PS on frequencies below 1 mm^{-1}, while quantum and electronic noise dominates in higher frequencies (KAVURI; DAS, 2020). To better visualize the contributions of quantum and electronic noise, we also show the zoom on these high frequencies in Fig. 33 (b). First, we can see that the degradation caused by spatial noise correlation decreases the high-frequency noise and slightly increases the low-frequency noise. This can be seen when comparing the blue and orange curves from the intersection point at frequency $2.5mm^{-1}$. For the network result, if we compare the PS of the input image with the restored one, it is possible to verify that the network adjusted the spectral dependence of the noise so that the output image had the PS very close to the target image (acquired with independent noise at 90mAs). However, by analyzing an ROI of the restored image (Fig. 31 (b)), it is possible to verify that the network filtered the low-frequency noise but added high-frequency noise to adjust the PS according to the target image (Fig. 31(c)). This high-frequency PS difference is



(a) (b)

Figure 33 – (a) Power Spectra (PS) calculated at the testing step of the proposed network cGAN #1, where the network performs only noise decorrelation; (b) zoom-in on PS high-frequencies where quantum and electronic noise is known to predominate.



(a) (b)

Figure 34 – (a) Power Spectra (PS) calculated at the testing step of the proposed network cGAN #2, where the network performs both denoising and decorrelation; (b) zoom-in on PS high-frequencies where quantum and electronic noise is known to predominate

illustrated by the two-head read arrow on Fig. 33 (b). We also plotted the PS for the GT for reference. Note that this image does not contain any noise.

The second approach (cGAN #2) was to target the green PS curve of Fig. 34, where the network task is to perform both denoising and decorrelation. This approach aims to mitigate the high-frequency noise injection observed in the first network. Aiming for an image with lower high-frequency noise, as observed in the PS curves, we constrain the network to not inject high-frequency white noise with a higher dose target (green line).

With that, the network performs both denoising and decorrelation successfully. We can see that, comparing the green and brown lines of Fig. 34, the network approximates well the PS, and through Fig. 31 (e) is possible to note that both denoising and decorrelation were done reasonably well.

Table 18 – Mean value of the MNSE calculated for all DBT projections of the synthetic breast phantom used for testing. The confidence interval is displayed inside the brackets. Input means the original correlated 90mAs image, output #1 the result cGan for 90mAs images, output #2 the cGan result for 360mAs images, target #1 the white noise 90mAs image and target #2 the white noise 360mAs image.

	Total MNSE(%)	$\mathcal{R}_{\mathcal{N}}(\%)$	$\mathcal{B}^2(\%)$
Input	37.50 [37.46, 37.55]	37.50 [37.49, 37.51]	0.00 [0.00, 0.00]
Output #1	37.29 [37.25, 37.34]	36.07 [36.06, 36.08]]	1.22 [1.22, 1.23]
Output #2	17.67 [17.63, 17.72]	15.02 [15.01, 15.02]	2.66 [2.65, 2.66]
Target #1	37.57 [37.55, 37.59]	37.57 [37.56, 37.58]	0.00 [0.00, 0.00]
Target #2	9.23 [9.22, 9.23]	9.23 [9.23, 9.23]	0.00 [0.00, 0.00]

Table 18 shows the image fidelity metrics for the testing phantom. The first network (cGan #1) matches the $\mathcal{R}_{\mathcal{N}}$ of the target but generated a $\mathcal{B}^2$ of 1.22%. The bias might be explained by the low-frequency denoising and the injection of high-frequency noise. For the second network (cGan #2), the restoration reduced the $\mathcal{R}_{\mathcal{N}}$ of the input image from 37.5% to 15.02% and generated a $\mathcal{B}^2$ of 2.66%. Note, however, that for this case, the network performed a selective filtering, based on the spatial frequency of the noise, *i.e.*, low-frequency noise was removed with more intensity than the high-frequency noise, to generate an output image with approximately no-correlated residual noise (white).

6.5 Conclusions

In conclusion, CNNs might be used to restore images corrupted by correlated noise. We found that the cycle-GAN was able to perform both denoising and decorrelation. The network generated restored images with noise frequency components very close to the target images. This costs an increase in the bias of the output image. Further studies might be done in the sense of imposing image fidelity to the network to obtain lower bias values while achieving satisfactory noise frequency properties. For the clinical results, it was demonstrated that the network did not overfit to the simulated dataset, since it achieved good results on the testing dataset. To further improve the results, especially for microcalcifications, we can insert real calcifications on the simulated dataset and train the network so it learns how to preserve more these fine details.

Acknowledgments

This work was supported in part by the São Paulo Research Foundation (FAPESP grant 2021/12673-6 and 2022/04646-1), by the National Council for Scientific and Technological Development (CNPq) and by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001).

7 PAPER 5:COMPUTATIONAL METHOD TO ARTIFICIALLY INSERT CLUSTERS OF MICROCALCIFICATIONS IN DIGITAL BREAST TOMOSYNTHESIS

The material presented in this chapter was published in: Vimieiro, Rodrigo de Barros, Lucas Rodrigues Borges, Bruno Barufaldi, Andrew Douglas Arnold Maidment, Renato França Caron, Silvia Maria Prioli De Souza Sabino, Ge Wang, and Marcelo Andrade da Costa Vieira. "Computational method to artificially insert clusters of microcalcifications in digital breast tomosynthesis." In Medical Imaging 2023: Physics of Medical Imaging, vol. 12463, p. 572-578. SPIE, 2023. Available at: https://doi.org/10.1117/12.2653857. The author obtained permission for total and/or partial reproduction of the content from the publisher SPIE.

7.1 Abstract

Digital Breast Tomosynthesis (DBT) is a medical imaging modality that has been increasingly used for breast cancer screening. To improve the accuracy in the early detection of breast cancer, it is common to use tools based on image processing to improve the quality of DBT images and, consequently, the visibility of lesions of clinical interest. Microcalcification (MC) clusters are in the class of findings that may indicate the early stages of breast cancer. To evaluate the impact of image processing methods in the detection of breast lesions, human perception studies are usually performed, where detection accuracy is evaluated using positive and negative cases, i.e., images with and without lesions. However, only a small fraction of clinical cases have positive cases and the exact location of the lesion is not always known. Thus, in this work, we present a method to artificially insert MC cluster in normal exams of DBT. A set of MCs was segmented from clinical cases acquired in a prone stereotactic breast biopsy system. In the pipeline, we built a MC cluster and insert it in a 3D volume. Then, we project it on the detector and inserted the MC cluster into the projections of a clinical DBT exam. It is possible to define the size, contrast, and location of the MC cluster accordingly. We performed a two-alternative forced-choice (2-AFC) study with six experienced medical physicists and the average success rate was 53.83%, suggesting that readers, on average, could not distinguish between real and simulated MCs. Overall results showed that images with simulated MC are similar to the real clinical cases. Our source code is available at www.github.com/LAVI-USP/MCInsertionPackage-DBT.

7.2 Introduction

The development and validation of many medical image processing algorithms rely on the variability of the data available. Among the medical imaging fields that greatly benefit from data variability, we can mention learning-based algorithms trained to classify, segment, and restore medical images. One of the challenges of creating a diverse dataset in medical imaging is to identify abnormalities, as they may be infrequent or difficult to detect. One approach to overcome this limitation is to simulate the abnormalities with the desired characteristics (e.g., intensity, size, localization) and later embed the simulated abnormality into healthy clinical patient data, thus creating a hybrid abnormal case. In mammography, microcalcification (MC) clusters are in the class of abnormalities that may sign breast cancer (YAFFE, 2000), and the virtual insertion of MCs on healthy exams has been done in previous works (SHANKLA *et al.*, 2014; SHAHEEN *et al.*, 2011; SHAHEEN *et al.*, 2010; BORGES; MARQUES; VIEIRA, 2019).

Recently, Borges *et al.* (BORGES; MARQUES; VIEIRA, 2019) developed a method to artificially insert MC clusters into full-field digital mammography (FFDM) exams. This work extracted real MCs from mammography exams followed by the insertion of the lesions into normal exams, taking into consideration their localization, contrast and size. The method was previously used by our group to evaluate the effect of denoising methods on the localization of MCs (BORGES *et al.*, 2020). The objective of this work is to further expand our previous work for Digital Breast Tomosynthesis (DBT), allowing the artificial insertion of MCs into DBT exams with no findings.

7.3 Materials & methods

7.3.1 Simulation pipeline

The pipeline proposed in this work is summarized in Fig. 35. In the first step, lesion-free DBT projections were identified from a dataset of retrospective exams, as described in section 7.3.3. The publicly available[1] Laboratory for Breast Radiodensity Assessment (LIBRA) software (KELLER *et al.*, 2012; KELLER *et al.*, 2015) was used in the next step to estimate breast density projection masks. We used morphological operations to erode the masks and remove isolated points. The estimated density masks were back-projected using our in-house DBT reconstruction toolbox (VIMIEIRO; BORGES; VIEIRA, 2019), publicly available for download[2]. To identify areas of dense tissue in the back-projected masks we adopted a hard-threshold of 0.5 pixels, with values above 0.5 labeled as candidate locations for the MC cluster. Finally, one of the candidate locations of the cluster was randomly selected.

The simulated MC clusters were created using individual MC from the dataset described in Maidment *et al.* (MAIDMENT *et al.*, 1996), spatially arranged using a stack of 2D Gaussian probability density functions - a 3D generalization of the process described in Borges *et al.* (BORGES; MARQUES; VIEIRA, 2019). The number of individual MC

[1] https://www.med.upenn.edu/sbia/libra.html
[2] https://github.com/LAVI-USP/pyDBT

per cluster was randomly defined using a uniform distribution with probabilities at the range 1 to 4. After the MC cluster was generated, we used our in-house reconstruction toolbox (VIMIEIRO; BORGES; VIEIRA, 2019) to project the lesion volume into 2D projections, using the appropriate geometry. Each individual MC was simulated with a different contrast, similar to Borges *et al.* (BORGES; MARQUES; VIEIRA, 2019). This was done by projecting each MC individually and fixing the contrast in the projection domain according to the size of that MC. After the projection of all MCs, we applied the DBT system modulation transfer function (MTF) to the images. We estimated the Senographe Pristina (GE) MTF following the procedure presented on the Equipment Report Technical evaluation (ENGLAND, 2019). Finally, we embedded the MC clusters into the clinical images using an overall contrast of 0.2, defined empirically. Note that the code is flexible and allows the modification of any parameter, so the simulation may be tuned to reflect a particular population.

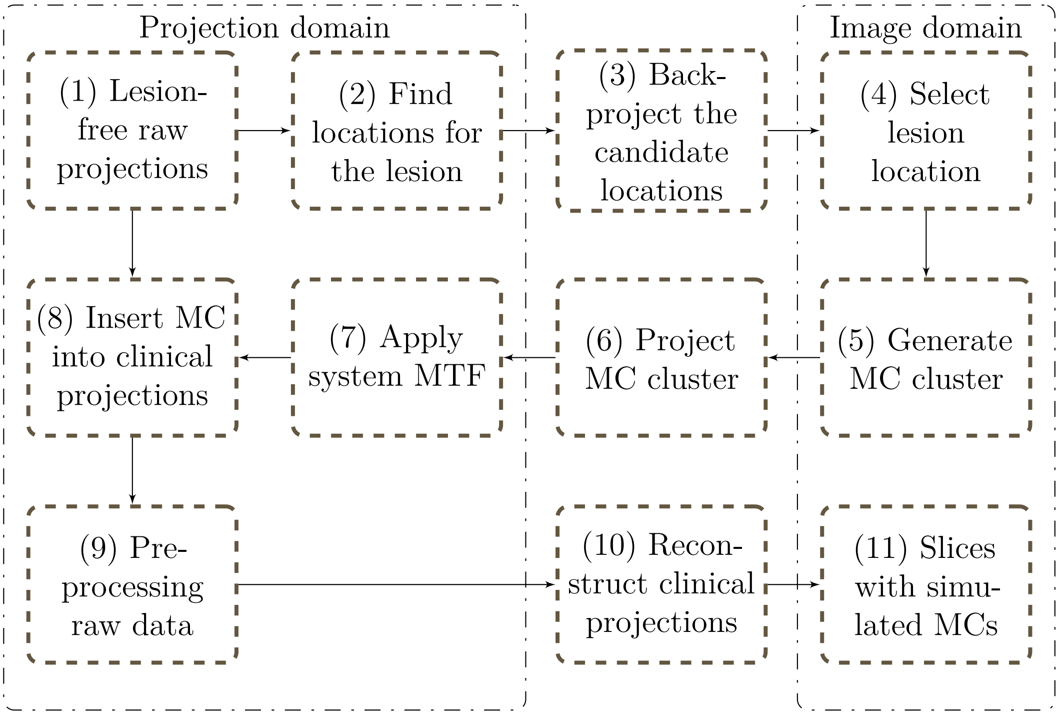

Projection domain
Image domain
(1) Lesion-free raw projections
(2) Find locations for the lesion
(3) Back-project the candidate locations
(4) Select lesion location
(5) Generate MC cluster
(6) Project MC cluster
(7) Apply system MTF
(8) Insert MC into clinical projections
(9) Pre-processing raw data
(10) Reconstruct clinical projections
(11) Slices with simulated MCs

Figure 35 – Scheme illustrating the proposed pipeline to generate and insert MC cluster in DBT projections.

7.3.2 Two-alternative forced-choice

To assess the appearance of the simulated clusters, we performed a two-alternative forced-choice (2-AFC) study with six medical physicists with an average of five years of experience in breast imaging. A total of 100 image pairs were displayed and each pair contained ROIs of one real and one simulated MC cluster. The region of interest had the size of $20 \times 20 \times 10mm$ and the user was able to scroll through 15 slices. We created a graphic user interface[3] where the reader was asked which images is the real one, as illustrated on Fig. 36.

A correct selection rate close to 50% may indicate that the readers were not able to distinguish between real and simulated MC clusters, while rates above (or below) 50% may indicate that the readers were able to identify differences.

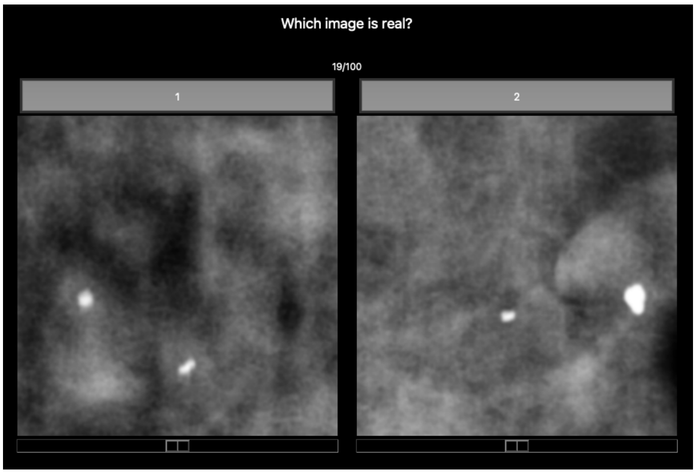


Figure 36 – The graphic user interface used by readers to perform the two-alternative forced-choice test. One real and one simulated MC cluster are displayed.

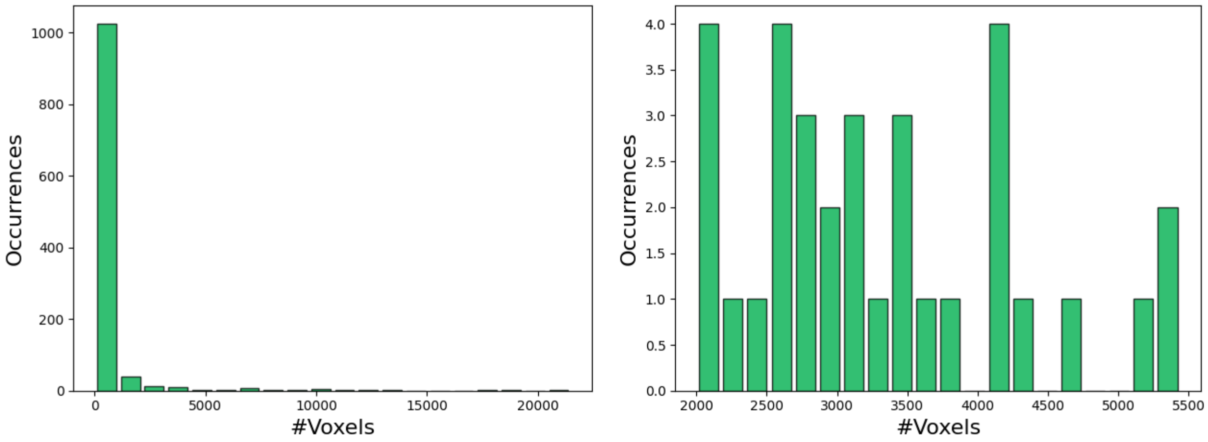
7.3.3 Lesion-free and lesion-present dataset

Retrospective exams from 44 patients were inspected by two experienced medical physicists and 64 lesion-free DBT exams were identified. A total of 100 MC clusters were simulated and inserted into the lesion-free cases. Note that 36 samples had 2 inserted MC clusters. However, since we are using ROIs in our human-observer test, different regions were selected for those exams. A separate cohort of 100 patients, with identified MC clusters were also used in the 2-AFC study (benchmark dataset). All the clinical samples used in this study were obtained in a collaboration with the Barretos Cancer Hospital (Brazil) and were acquired using a Senographe Pristina (GE) DBT System under IRB approval *CAAE #986670*.

7.3.4 Microcalcification dataset

Individual calcifications were selected from the dataset published in Maidment *et al.* (MAIDMENT *et al.*, 1996). Although the dataset consists of both individual MCs and clusters of MCs, in this work we only used individual MCs. There are 1,117 individual MCs and the distribution in terms of volume is shown in Fig. 37 (a).

[3] https://github.com/LAVI-USP/2AFC_Interface



(a) (b)

Figure 37 – (a) Distribution of volume, in number of voxels, of MCs for the entire dataset. (b) Distribution of volume, in number of voxels, of MCs, after the selection criteria.

To match the characteristics of our cohort of lesion-present dataset, we selected individual MCs with the number of voxels between 2000 and 6000. Figure 37 (b) illustrates the distribution of the volume after the selection criteria.

7.3.5 Image reconstruction

After simulating the MC cluster for 100 cases, we back-projected the volume using our in-house toolbox. The raw projections were processed using the commercially available software Briona Standards (Real Time Tomography, PA, USA)[4]. The same procedure was done to reconstruct the 100 case with lesion present (benchmark).

7.4 Results & discussions

7.4.1 Simulated images

Figure 38 illustrates ROIs with real and simulated MC. The first row shows real MCs extracted from the dataset with MC clusters. The second row shows some simulated MC cases. Note that these ROIs were extracted from DBT slices. Overall, it is possible to say that the simulated MC morphology is very similar to the real ones. Note that the variability of shape, contrast and size of the simulated MCs approximate to those from the real clusters.

Figure 39 shows a sequence of 4 slices from a simulated MC. For the sake of simplicity we are just showing 4 slices out of 15. As we mentioned before, each calcification is inserted at a random coordinate for X,Y and Z. In this figure we can see that on slice #1 both are out of focus. This also happens on the slice #15. However, in slice #5 the

[4] https://www.realtimetomography.com

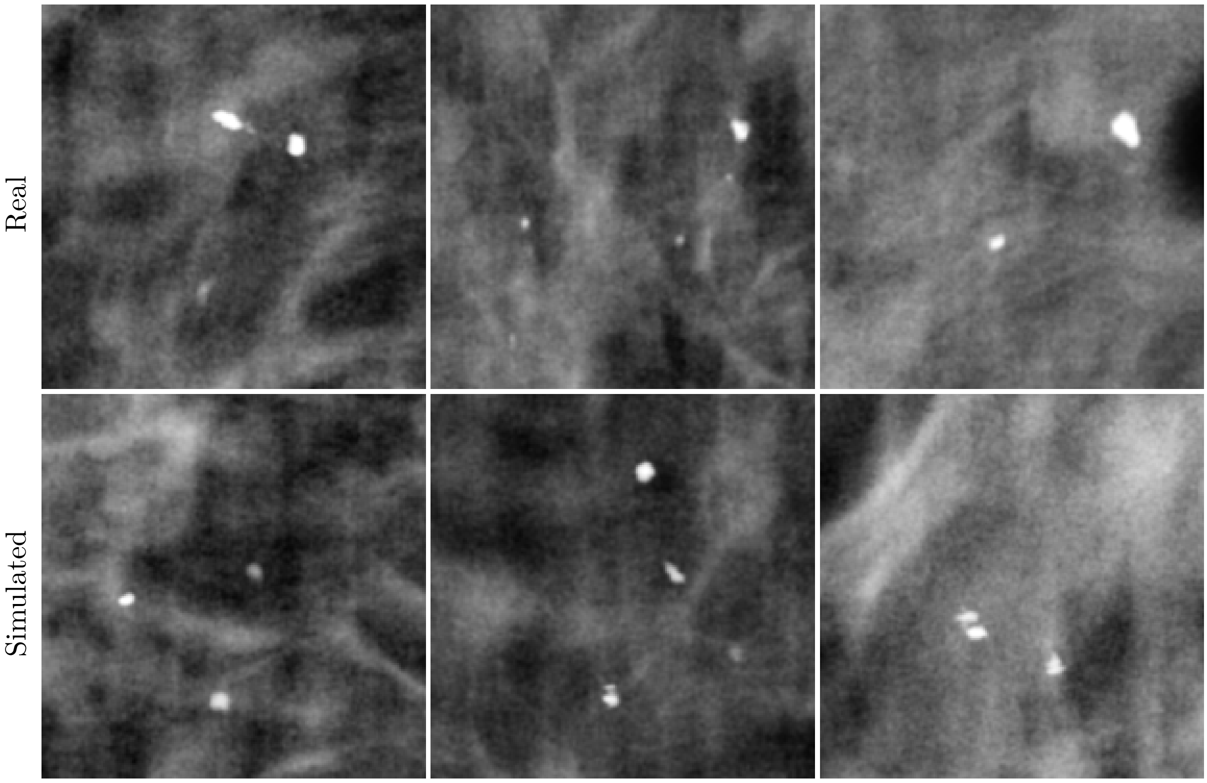
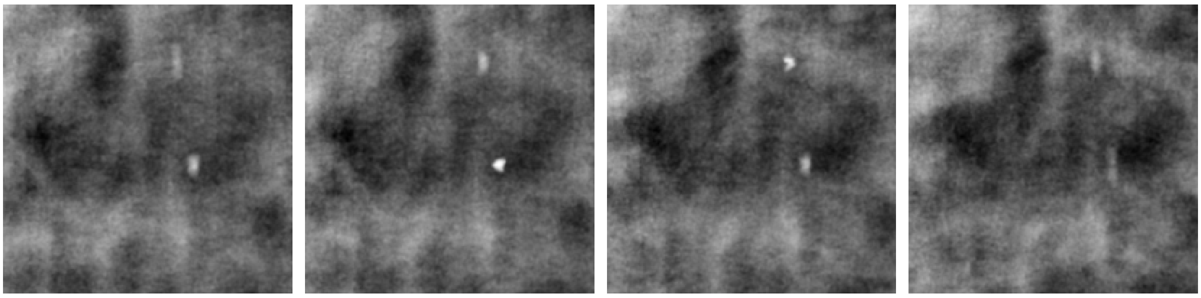


Figure 38 – Illustration of real (first row) and simulated MCs (second row). The ROIs were extracted from DBT slices.

bottom calcification is on focus while the top one is out of focus and the inverse happens on slice #10.



(a) slice #01 (b) slice #05 (c) slice #10 (d) slice #15

Figure 39 – Illustration of 4 slices out of the 15 from the simulated cluster. The figure shows simulated MCs on different depths.

7.4.2 Two-alternative forced-choice study

The 2-AFC results are shown in Fig. 40. As previously mentioned, it is expected that readers get the correct selection rate close to 50% if they are not able to identify differences between real and simulated clusters. The gray region shows the 95% CI of random choice [39.8% 60.1%] for one reader (N=100). Considering the entire test (N=600), the 95% confidence interval (CI) is [45.9% 54.0%], according to the Binomial distribution.

The mean successful rate of the study was 53.83%. The mean successful rate is within the CI of the study suggesting that, on average, the readers could not distinguish between the real and simulated cases.

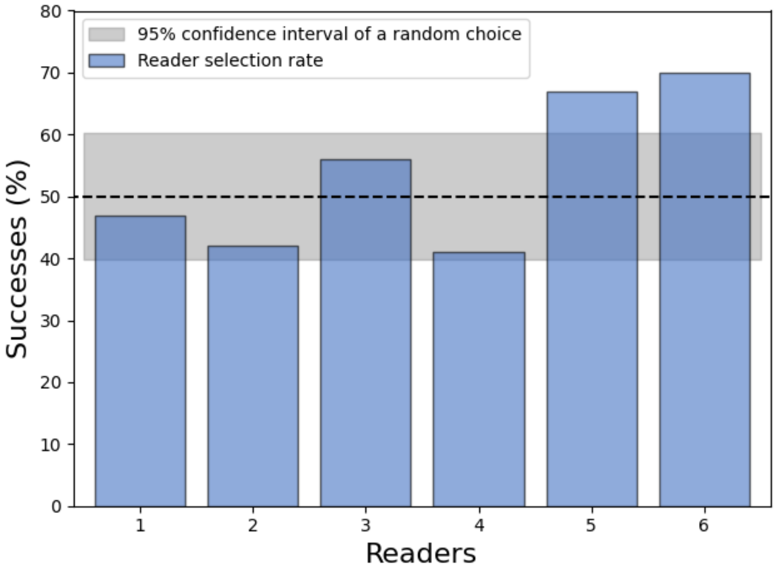


Figure 40 – 2-AFC results for each individual reader. The gray region shows the confidence interval for the random selection of a single reader (Binomial distribution with N=100).

The results show that 4 out of 6 readers had the selection rate within the 95% CI of random selection. Two readers (reader #5 and #6) tended to choose the real MCs. When questioned about what helped them to identify the correct lesions, reader #5 stated that he identified that in all simulated cases the center of the MC's volume of interest matched the central slice of the volume of interest (VOI) being displayed. Reader #6 stated that he noticed more out-of-plane artifacts in simulated cases if compared to real ones. In real cases, sometimes the center of the MC's VOI was slightly shifted with respect to the center of the displayed stack. This is a limitation of this study, as the center of the simulated VOI is known exactly while the center of the real VOI can only be estimated by the reader who identified the lesion.

7.5 Conclusions

In this work we presented a method to simulate MC clusters in DBT exams. We inserted MCs in the image domain with different contrasts, in dense regions and projected them onto the detector. To verify the aspect of the simulated lesions, we performed a 2-AFC study with six experienced medical physicists. The average selection rate was within the CI of random selection suggesting that readers, on average, could not differentiate between the simulated and real MC clusters. On the individual level, two readers were better at detecting the simulated clusters.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP) grant #2021/12673-6, by the National Council for Scientific and Technological Development (CNPq), by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES finance code 001) and by the National Institute of Health (R01EB026646, R01CA233888, R01CA237267, and R01HL151561). The authors would like to thank Real Time Tomography for providing access to the image processing software and the team of medical physicists who volunteered to participate in the 2-AFC experiment.

8 CONCLUSIONS

To investigate MBDL approaches, we first start developing DB methods and all their characteristics. More specifically, in Paper 1, we observed that NNs have the potential to restore LD FFDM images, however, their results are slightly worse compared to the MB one. Obviously, this work has its limitations since it uses a simple architecture. With more advanced architectures it is expected that the results would improve. We also observed the potential of each loss function and observed that content losses like MAE and MSE impose blur but tend to keep image fidelity. On the other hand, visual perception losses can be used to retrieve image details and decrease the signal blur. In conclusion, we showed the potential of DL models for medical imaging restoration and the great importance of the loss function on this task. Since ML methods in the CV area do not only depend on the loss, Paper 2 showed different training strategies for image restoration. In conclusion, different training strategies such as early stops can be done depending on the target task and we also showed the importance of cross-validation. In Paper 3, a loss function considering the power spectrum of the images was proposed. This necessity comes from the fact that visual perception losses like the PL impose spatial noise correlation in their results. We observed that the combination with the proposed loss was able to keep the power spectrum properties while maintaining the image details.

In Paper 4, the synergy of MB methods and DB approaches have great potential to be explored in the DL field, especially in image restoration. Since DL methods suffer from black-box nature, introducing known and established mathematical formulations on the architecture design, improves not only the results but also the stability of the models. Over the initial works, we built knowledge on DB models that learn specifically from the provided data and afterward, we compared this model with the proposed MBDL framework. We observed the black-box nature of solely DB approaches in Paper 2, where cross-validation showed different objective results each time. However, in the MBDL approach, even with different architecture, they resulted in very similar metrics, which suggest more stability and control over the models. We demonstrated that the networks benefit from all this previous knowledge added, independent of the utilized NN architecture. It is still not clear if a final image with slightly high RN and lower bias, compared to the target FD, is better compared to another one with lower RN but higher bias. This question must be answered with clinical and human perception studies. Independent of the answer, the proposed bias-RN loss function is able to set the network at some desired point of operation in terms of these two metrics. Obviously, this point is limited to the performance of the chosen architecture, as we noticed in the experiments. However, the whole proposed framework is architecture-agnostic, making it suitable for different ones and perhaps more advanced ones.

In paper 5, we demonstrated the potential of DNNs to restore correlated noisy images, which is different from the independent noise properties from the first papers. It was able to both denoise and decorrelate the degraded images. However, this work has limitations. We only used one architecture inside the cycle-GAN framework. Also, it is still not clear whether the bias imposed in the process might impact localization and cancer detection or not. In conclusion, the use of a precise model to simulate the noise correlation kernel and apply it to the images, with this specific degradation, imposes some indirect prior knowledge on the models, making it possible to have target images for supervised learning. In real scenarios, such an approach would not be possible. Although further validation is necessary, preliminary results indicate that the MBDL may be suitable for this task.

Finally, in Paper 6 we proposed a method to simulate MC on DBT images. This is the first step to performing clinical studies for Paper 5. In conclusion, the 2-AFC test showed, on average, that the simulated MC could not be distinguished from the real ones. This work was limited since it was only evaluated by medical physicists and for future work it should be tested with radiologists.

For future work, in the MBDL framework, we can explore the potential of using a learnable weight in the last residual layer. Theoretically, the mathematical model considers that the denoising algorithm reduces the noise variance to zero. This is not always true in reality. So, a learnable parameter could compensate for this limitation. Also, different stabilization transforms can be tested and perhaps end-to-end training can be done, eliminating the necessity of using a self-learning framework. Several state-of-the-art architectures can be tested to improve the objective metrics even more.

It is important also to test the framework in systems with frequency-dependent correlated noise. As a first step, we used the cycle-GAN to restore this type of image. Combining this approach with the MBDL framework might improve the results even more. Another potential approach would be combining the residual sum used in the MBDL framework to restore image details for the four-fold restoration. We could use the one-fold uncorrelated image to retrieve some details of the four-fold restoration, *i.e.*, a combination of both approaches proposed in Paper 5 with the framework of Paper 3. Indeed, this denoising with decorrelation must be tested in practice with human perception studies.

As a general conclusion, the synergy of MB methods and DB approaches has great potential to be explored within the DL domain, demonstrated by the improved results over models that do not benefit from those priors. All these AI methods used in the field of medical imaging have the potential of improving image quality, reducing radiation dose and perhaps enhancing breast cancer diagnosis.

REFERENCES

ADLER, J.; ÖKTEM, O. Learned primal-dual reconstruction. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1322–1332, 2018.

AL-MASNI, M. A. *et al.* Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 157, p. 85–94, 2018.

ARAÚJO, D. M. de; SALVADEO, D. H.; PAULA, D. D. de. Denoising digital breast tomosynthesis projections using convolutional neural networks. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2021: Image Processing**. [*S.l.: s.n.*], 2021. v. 11596, p. 115961L.

AREFAN, D. *et al.* Deep learning modeling using normal mammograms for predicting breast cancer risk. **Medical Physics**, Wiley Online Library, v. 47, n. 1, p. 110–118, 2020.

AZZARI, L.; BORGES, L. R.; FOI, A. Chapter 1 - Modeling and Estimation of Signal-Dependent and Correlated Noise. *In*: BERTALMÍO, M. (ed.). **Denoising of Photographic Images and Video: Fundamentals, Open Challenges and New Trends**. Switzerland: Springer, 2018. p. 13–36.

BAKIC, P. R. *et al.* **The open-source virtual clinical trial project**. 2017. (Accessed: 16 Jun 2021), https://sourceforge.net/projects/openvct.

BAKIC, P. R. *et al.* Testing realism of software breast phantoms: Texture analysis of synthetic mammograms. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2013: Physics of Medical Imaging**. [*S.l.: s.n.*], 2013. v. 8668, p. 866824.

BARUFALDI, B. *et al.* OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2018: Physics of Medical Imaging**. [*S.l.: s.n.*], 2018. v. 10573, p. 1057358.

BATSON, J.; ROYER, L. Noise2self: Blind denoising by self-supervision. *In*: PMLR. **International Conference on Machine Learning**. [*S.l.: s.n.*], 2019. p. 524–533.

BOITA, J. *et al.* How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography? **European Radiology**, Springer, p. 1–9, 2021.

BORGES, L. R. *et al.* Restoration of low-dose digital breast tomosynthesis. **Measurement Science and Technology**, IOP Publishing, v. 29, n. 6, p. 064003, 2018.

BORGES, L. R. *et al.* Pipeline for effective denoising of digital mammography and digital breast tomosynthesis. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2017: Physics of Medical Imaging**. [*S.l.: s.n.*], 2017. v. 10132, p. 1013206.

BORGES, L. R. *et al.* Technical note: Noise models for virtual clinical trials of digital breast tomosynthesis. **Medical Physics**, 2019.

BORGES, L. R. *et al.* Effect of denoising on the localization of microcalcification clusters in digital mammography. *In*: SPIE. **15th International Workshop on Breast Imaging (IWBI2020)**. [*S.l.: s.n.*], 2020. v. 11513, p. 149–155.

BORGES, L. R. *et al.* Method for simulating dose reduction in digital breast tomosynthesis. **IEEE Transactions on Medical Imaging**, IEEE, v. 36, n. 11, p. 2331–2342, 2017.

BORGES, L. R.; MARQUES, P. M. de A.; VIEIRA, M. A. A 2-AFC study to validate artificially inserted microcalcification clusters in digital mammography. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment**. [*S.l.: s.n.*], 2019. v. 10952, p. 109520R.

BORGES, L. R. *et al.* Method for simulating dose reduction in digital mammography using the Anscombe transformation. **Medical Physics**, Wiley Online Library, v. 43, n. 6Part1, p. 2704–2714, 2016.

BROWN, T. *et al.* Language models are few-shot learners. **Advances in Neural Information Processing Systems**, v. 33, p. 1877–1901, 2020.

CARTON, A.-K. *et al.* Development of a physical 3D anthropomorphic breast phantom. **Medical Physics**, Wiley Online Library, v. 38, n. 2, p. 891–896, 2011.

CHAN, H.-P. *et al.* Effect of dose level on radiologists' detection of microcalcifications in digital breast tomosynthesis: An observer study with breast phantoms. **Academic Radiology**, Elsevier, 2020.

CHAN, H.-P.; SAMALA, R. K.; HADJIISKI, L. M. CAD and AI for breast cancer—recent development and challenges. **The British Journal of Radiology**, The British Institute of Radiology., v. 92, p. 20190580, 2019.

CHEN, H. *et al.* LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1333–1347, 2018.

CHEN, H. *et al.* Low-dose CT with a residual encoder-decoder convolutional neural network. **IEEE Transactions on Medical Imaging**, IEEE, v. 36, n. 12, p. 2524–2535, 2017.

CHEN, H. *et al.* Low-dose CT via convolutional neural network. **Biomedical Optics Express**, Optical Society of America, v. 8, n. 2, p. 679–694, 2017.

CHEON, M. *et al.* Perceptual image quality assessment with transformers. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2021. p. 433–442.

CIOMPI, F. *et al.* Towards automatic pulmonary nodule management in lung cancer screening with deep learning. **Scientific Reports**, Nature Publishing Group, v. 7, p. 46479, 2017.

COSTA, A. C. *et al.* Transfer learning in deep convolutional neural networks for detection of architectural distortion in digital mammography. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **15th International Workshop on Breast Imaging (IWBI2020)**. [*S.l.: s.n.*], 2020. v. 11513, p. 115130N.

DABOV, K. *et al.* Image denoising by sparse 3-D transform-domain collaborative filtering. **IEEE Transactions on Image Processing**, IEEE, v. 16, n. 8, p. 2080–2095, 2007.

DAS, M. *et al.* Penalized maximum likelihood reconstruction for improved microcalcification detection in breast tomosynthesis. **IEEE Transactions on Medical Imaging**, IEEE, v. 30, n. 4, p. 904–914, 2010.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DING, K. *et al.* Comparison of full-reference image quality models for optimization of image processing systems. **International Journal of Computer Vision**, Springer, v. 129, n. 4, p. 1258–1281, 2021.

DOBBINS, J. T. Image quality metrics for digital systems. *In*: METTER, R. L. V.; BEUTEL, J.; KUNDEL, H. L. (ed.). **Handbook of Medical Imaging: Volume 1. Physics and Psychophysics**. 1. ed. Bellingham: SPIE Press, 2000. p. 163–219.

ENGLAND, P. H. **NHS Breast Screening Programme Equipment Report Technical evaluation of GE Healthcare Senographe Pristina digital breast tomosynthesis system**. [*S.l.: s.n.*]: Public Health England, 2019.

ERHAN, D. *et al.* **Visualizing higher-layer features of a deep network**. University of Montreal, 2009. v. 1341, n. 3, 1 p.

FRIEDEWALD, S. M. *et al.* Breast cancer screening using tomosynthesis in combination with digital mammography. **JAMA**, American Medical Association, v. 311, n. 24, p. 2499–2507, 2014.

GAO, M.; FESSLER, J. A.; CHAN, H.-P. Deep convolutional neural network with adversarial training for denoising digital breast tomosynthesis images. **IEEE Transactions on Medical Imaging**, IEEE, 2021.

GAO, M. *et al.* Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2020: Physics of Medical Imaging**. [*S.l.: s.n.*], 2020. v. 11312, p. 113120Q.

GARRETT, J. W. *et al.* Reduced anatomical clutter in digital breast tomosynthesis with statistical iterative reconstruction. **Medical Physics**, Wiley Online Library, v. 45, n. 5, p. 2009–2022, 2018.

GHETTI, C. *et al.* Physical characteristics of GE Senographe Essential and DS digital mammography detectors. **Medical Physics**, Wiley Online Library, v. 35, n. 2, p. 456–463, 2008.

GONG, K. *et al.* PET image reconstruction using deep image prior. **IEEE Transactions on Medical Imaging**, IEEE, v. 38, n. 7, p. 1655–1665, 2018.

GOODFELLOW, I. *et al.* Generative adversarial nets. *In*: **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*], 2014. p. 2672–2680.

GREEN, M. *et al.* Neural Denoising of Ultra-low Dose Mammography. *In*: SPRINGER. **International Workshop on Machine Learning for Medical Image Reconstruction**. [*S.l.: s.n.*], 2019. p. 215–225.

GU, J. *et al.* Recent advances in convolutional neural networks. **Pattern Recognition**, Elsevier, v. 77, p. 354–377, 2018.

GULRAJANI, I. *et al.* Improved training of Wasserstein GANs. **Advances in Neural Information Processing Systems**, v. 30, 2017.

HAN, M.; BAEK, J. Low-dose CT denoising via CNN with an observer loss function. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2021: Physics of Medical Imaging**. [*S.l.: s.n.*], 2021. v. 11595, p. 1159544.

HARVEY, H. *et al.* The role of deep learning in breast screening. **Current Breast Cancer Reports**, Springer, v. 11, n. 1, p. 17–22, 2019.

HAUS, A. G.; YAFFE, M. J. Screen-film and digital mammography: image quality and radiation dose considerations. **Radiologic Clinics of North America**, Elsevier, v. 38, n. 4, p. 871–898, 2000.

HE, K. *et al.* Deep residual learning for image recognition. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2016. p. 770–778.

HENDRIKSEN, A. A.; PELT, D. M.; BATENBURG, K. J. Noise2inverse: Self-supervised deep convolutional denoising for tomography. **IEEE Transactions on Computational Imaging**, IEEE, v. 6, p. 1320–1335, 2020.

HEUSEL, M. *et al.* GANs trained by a two time-scale update rule converge to a local nash equilibrium. **Advances in Neural Information Processing Systems**, v. 30, 2017.

HOOLEY, R. J.; DURAND, M. A.; PHILPOTTS, L. E. Advances in digital breast tomosynthesis. **American Journal of Roentgenology**, Am Roentgen Ray Soc, v. 208, n. 2, p. 256–266, 2017.

HUA, K.-L. *et al.* Computer-aided classification of lung nodules on computed tomography images via deep learning technique. **OncoTargets and Therapy**, Dove Press, v. 8, 2015.

HUANG, G. *et al.* Densely connected convolutional networks. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2017. p. 4700–4708.

HUDA, W. *et al.* Experimental investigation of the dose and image quality characteristics of a digital mammography imaging system. **Medical Physics**, Wiley Online Library, v. 30, n. 3, p. 442–448, 2003.

IAEA. **Radiation Protection and Safety in Medical Uses of Ionizing Radiation**. Vienna: International Atomic Energy Agency, 2018. (Specific Safety Guides, SSG-46). ISBN 978-92-0-101717-8.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. **arXiv preprint arXiv:1502.03167**, 2015.

JOHNSON, J.; ALAHI, A.; FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. *In*: SPRINGER. **European Conference on Computer Vision**. [*S.l.: s.n.*], 2016. p. 694–711.

JR, R. S. S. *et al.* Does image quality matter? Impact of resolution and noise on mammographic task performance. **Medical Physics**, Wiley Online Library, v. 34, n. 10, p. 3971–3981, 2007.

KALRA, M. K. *et al.* Low-dose CT of the abdomen: evaluation of image improvement with use of noise reduction filters—pilot study. **Radiology**, Radiological Society of North America, v. 228, n. 1, p. 251–256, 2003.

KANG, E. *et al.* Deep convolutional framelet denosing for low-dose CT via wavelet residual network. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1358–1369, 2018.

KANG, E.; MIN, J.; YE, J. C. A deep convolutional neural network using directional wavelets for low-dose x-ray CT reconstruction. **Medical Physics**, Wiley Online Library, v. 44, n. 10, p. e360–e375, 2017.

KAVURI, A.; DAS, M. Relative contributions of anatomical and quantum noise in signal detection and perception of tomographic digital breast images. **IEEE Transactions on Medical Imaging**, IEEE, v. 39, n. 11, p. 3321–3330, 2020.

KELLER, B. M. *et al.* Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography. **Breast Cancer Research**, Springer, v. 17, n. 1, p. 1–17, 2015.

KELLER, B. M. *et al.* Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. **Medical Physics**, Wiley Online Library, v. 39, n. 8, p. 4903–4917, 2012.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *In*: **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*], 2012. p. 1097–1105.

KRULL, A.; BUCHHOLZ, T.-O.; JUG, F. Noise2void-learning denoising from single noisy images. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2019. p. 2129–2137.

LAI, Y.-C. *et al.* Digital breast tomosynthesis: technique and common artifacts. **Journal of Breast Imaging**, Oxford University Press US, v. 2, n. 6, p. 615–628, 2020.

LANGLOTZ, C. P. *et al.* A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. **Radiology**, Radiological Society of North America, v. 291, n. 3, p. 781–791, 2019.

LEE, G.; FUJITA, H. **Deep learning in medical image analysis: challenges and applications**. [*S.l.: s.n.*]: Springer, 2020. v. 1213.

LEHTINEN, J. *et al.* Noise2noise: Learning image restoration without clean data. **arXiv preprint arXiv:1803.04189**, 2018.

LEVAKHINA, Y. *et al.* Weighted simultaneous algebraic reconstruction technique for tomosynthesis imaging of objects with high-attenuation features. **Medical Physics**, Wiley Online Library, v. 40, n. 3, 2013.

LI, Z. *et al.* Adaptive nonlocal means filtering based on local noise level for CT denoising. **Medical Physics**, Wiley Online Library, v. 41, n. 1, p. 011908, 2014.

LIU, J. *et al.* Radiation dose reduction in digital breast tomosynthesis (DBT) by means of deep-learning-based supervised image processing. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2018: Image Processing**. [*S.l.: s.n.*], 2018. v. 10574, p. 105740F.

MAIDMENT, A. D. *et al.* 3-D mammary calcification reconstruction from a limited number of views. *In*: SPIE. **Medical Imaging 1996: Physics of Medical Imaging**. [*S.l.: s.n.*], 1996. v. 2708.

MÄKINEN, Y.; AZZARI, L.; FOI, A. Collaborative filtering of correlated noise: Exact transform-domain variance for improved shrinkage and patch matching. **IEEE Transactions on Image Processing**, IEEE, v. 29, p. 8339–8354, 2020.

MAKITALO, M.; FOI, A. Optimal inversion of the generalized Anscombe transformation for poisson-gaussian noise. **IEEE Transactions on Image Processing**, IEEE, v. 22, n. 1, p. 91–103, 2012.

MAN, Q. D. *et al.* A two-dimensional feasibility study of deep learning-based feature detection and characterization directly from CT sinograms. **Medical Physics**, Wiley Online Library, v. 46, n. 12, p. e790–e800, 2019.

MANDUCA, A. *et al.* Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. **Medical Physics**, Wiley Online Library, v. 36, n. 11, p. 4911–4919, 2009.

MARROCCO, C. *et al.* Mammogram denoising to improve the calcification detection performance of convolutional nets. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **14th International Workshop on Breast Imaging (IWBI 2018)**. [*S.l.: s.n.*], 2018. v. 10718, p. 107180W.

MATHIASEN, A.; HVILSHØJ, F. Backpropagating through Fréchet Inception Distance. **arXiv preprint arXiv:2009.14075**, 2020.

MCCOLLOUGH, C. H. *et al.* Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge. **Medical Physics**, Wiley Online Library, v. 44, n. 10, p. e339–e352, 2017.

MCENTEE, M. F. Clinical radiographic units. *In*: RUSSO, P. (ed.). **Handbook of X-ray Imaging: Physics and Technology**. 1. ed. Boca Raton: CRC Press, 2017. cap. 26, p. 518 – 544.

MCKINNEY, S. M. *et al.* International evaluation of an AI system for breast cancer screening. **Nature**, Nature Publishing Group, v. 577, n. 7788, p. 89–94, 2020.

MICHELL, M.; BATOHI, B. Role of tomosynthesis in breast imaging going forward. **Clinical Radiology**, Elsevier, v. 73, n. 4, p. 358–371, 2018.

MORAN, N. *et al.* Noisier2noise: Learning to denoise from unpaired noisy data. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2020. p. 12064–12072.

NAGARE, M. *et al.* A Bias-Reducing Loss Function for CT Image Denoising. *In*: IEEE. **ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [*S.l.: s.n.*], 2021. p. 1175–1179.

NIU, C. *et al.* Noise suppression with similarity-based self-supervised deep learning. **IEEE Transactions on Medical Imaging**, IEEE, 2022.

PASZKE, A. *et al.* **Automatic differentiation in pytorch**. 2017. Available at: https://pytorch.org.

PRABHAT, K. *et al.* Deep neural networks-based denoising models for CT imaging and their efficacy. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2021: Physics of Medical Imaging**. [*S.l.: s.n.*], 2021. v. 11595, p. 115950H.

RADFORD, A. *et al.* **Improving language understanding by generative pre-training**. [*S.l.: s.n.*]: OpenAI, 2018.

RADFORD, A. *et al.* **Language models are unsupervised multitask learners**. [*S.l.: s.n.*]: OpenAI, 2019.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine learning in medicine. **New England Journal of Medicine**, Mass Medical Soc, v. 380, n. 14, p. 1347–1358, 2019.

RAN, M. *et al.* Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network. **Medical Image Analysis**, Elsevier, v. 55, p. 165–180, 2019.

RAVISHANKAR, S.; YE, J. C.; FESSLER, J. A. Image reconstruction: From sparsity to data-adaptive methods and machine learning. **Proceedings of the IEEE**, IEEE, v. 108, n. 1, p. 86–109, 2019.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *In*: SPRINGER. **International Conference on Medical Image Computing and Computer-assisted Intervention**. [*S.l.: s.n.*], 2015. p. 234–241.

SAADATMAND, S. *et al.* Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. **BMJ**, British Medical Journal Publishing Group, v. 351, p. h4901, 2015.

SAHARIA, C. *et al.* Image super-resolution via iterative refinement. **arXiv preprint arXiv:2104.07636**, 2021.

SAHU, P. *et al.* Using virtual digital breast tomosynthesis for de-noising of low-dose projection images. *In*: IEEE. **2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)**. [*S.l.: s.n.*], 2019. p. 1647–1651.

SALVAGNINI, E. *et al.* Effective detective quantum efficiency for two mammography systems: measurement and comparison against established metrics. **Medical Physics**, Wiley Online Library, v. 40, n. 10, p. 101916, 2013.

SAMALA, R. K. *et al.* Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. **IEEE Transactions on Medical Imaging**, IEEE, v. 38, n. 3, p. 686–696, 2018.

SHAHEEN, E. *et al.* Realistic simulation of microcalcifications in breast tomosynthesis. *In*: SPRINGER. **International Workshop on Digital Mammography**. [*S.l.: s.n.*], 2010. p. 235–242.

SHAHEEN, E. *et al.* The simulation of 3D microcalcification clusters in 2D digital mammography and breast tomosynthesis. **Medical Physics**, Wiley Online Library, v. 38, n. 12, p. 6659–6671, 2011.

SHAN, H. *et al.* Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. **Nature Machine Intelligence**, Nature Publishing Group, v. 1, n. 6, p. 269–276, 2019.

SHAN, H. *et al.* Impact of loss functions on the performance of a deep neural network designed to restore low-dose digital mammography. **Artificial Intelligence in Medicine**, Elsevier, v. 142, p. 102555, 2023.

SHAN, H. *et al.* 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1522–1534, 2018.

SHANKLA, V. *et al.* Automatic insertion of simulated microcalcification clusters in a software breast phantom. *In*: SPIE. **Medical Imaging 2014: Physics of Medical Imaging**. [*S.l.: s.n.*], 2014. v. 9033, p. 547–558.

SHEIKH, H. R.; BOVIK, A. C. Image information and visual quality. **IEEE Transactions on Image Processing**, IEEE, v. 15, n. 2, p. 430–444, 2006.

SHEIKH, H. R.; BOVIK, A. C.; VECIANA, G. D. An information fidelity criterion for image quality assessment using natural scene statistics. **IEEE Transactions on Image Processing**, IEEE, v. 14, n. 12, p. 2117–2128, 2005.

SHEN, L. *et al.* Deep learning to improve breast cancer detection on screening mammography. **Scientific Reports**, Nature Publishing Group, v. 9, n. 1, p. 1–12, 2019.

SHEN, R. *et al.* Multi-context multi-task learning networks for mass detection in mammogram. **Medical Physics**, Wiley Online Library, 2019.

SHLEZINGER, N. *et al.* Model-based deep learning. **Proceedings of the IEEE**, IEEE, 2023.

SILVER, D. *et al.* Mastering the game of go without human knowledge. **Nature**, Nature Publishing Group, v. 550, n. 7676, p. 354–359, 2017.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SKAANE, P. *et al.* Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. **Radiology**, Radiological Society of North America, Inc., v. 267, n. 1, p. 47–56, 2013.

STARCK, J.-L.; MURTAGH, F. D.; BIJAOUI, A. **Image processing and data analysis: the multiscale approach**. [*S.l.: s.n.*]: Cambridge University Press, 1998.

SUN, C. *et al.* Revisiting unreasonable effectiveness of data in deep learning era. *In*: **Proceedings of the IEEE International Conference on Computer Vision**. [*S.l.: s.n.*], 2017. p. 843–852.

SZEGEDY, C. *et al.* Going deeper with convolutions. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2015. p. 1–9.

TIAN, C. *et al.* Deep learning on image denoising: An overview. **Neural Networks**, Elsevier, 2020.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. **Nature Medicine**, Nature Publishing Group, v. 25, n. 1, p. 44–56, 2019.

VEDANTHAM, S. *et al.* Digital breast tomosynthesis: state of the art. **Radiology**, Radiological Society of North America, v. 277, n. 3, p. 663–684, 2015.

VIMIEIRO, R. B. *et al.* Assessment of training strategies for convolutional neural network to restore low-dose digital breast tomosynthesis projections. *In*: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2022: Physics of Medical Imaging**. SPIE, 2022. v. 12031. Available at: https://doi.org/10.1117/12.2609945.

VIMIEIRO, R. B.; BORGES, L. R.; VIEIRA, M. A. Open-source reconstruction toolbox for digital breast tomosynthesis. *In*: SPRINGER. **XXVI Brazilian Congress on Biomedical Engineering: CBEB 2018, Armação de Buzios, RJ, Brazil, 21-25 October 2018 (Vol. 2)**. [*S.l.: s.n.*], 2019. p. 349–354.

VIMIEIRO, R. B. *et al.* Convolutional neural network to restore low-dose digital breast tomosynthesis projections in a variance stabilization domain. **arXiv preprint arXiv:2203.11722**, 2022.

WALT, S. van der *et al.* scikit-image: image processing in python. **PeerJ**, v. 2, p. e453, jun 2014. ISSN 2167-8359. Available at: https://doi.org/10.7717/peerj.453.

WANG, G.; YE, J. C.; MAN, B. D. Deep learning for tomographic image reconstruction. **Nature Machine Intelligence**, Nature Publishing Group, v. 2, n. 12, p. 737–748, 2020.

WANG, G. *et al.* Image reconstruction is a new frontier of machine learning. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1289–1296, 2018.

WANG, J. *et al.* Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography. **IEEE Transactions on Medical Imaging**, IEEE, v. 25, n. 10, p. 1272–1283, 2006.

WANG, Z. *et al.* Image quality assessment: from error visibility to structural similarity. **IEEE Transactions on Image Processing**, IEEE, v. 13, n. 4, p. 600–612, 2004.

WANG, Z.; SIMONCELLI, E. P. Translation insensitive image similarity in complex Wavelet domain. *In*: IEEE. **Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005**. [*S.l.: s.n.*], 2005. v. 2, p. ii–573.

WHO. **World Health Organization - Breast Cancer**. 2023. www.who.int/. Access at: March 28 of 2023.

WOLTERINK, J. M. *et al.* Generative adversarial networks for noise reduction in low-dose CT. **IEEE Transactions on Medical Imaging**, IEEE, v. 36, n. 12, p. 2536–2545, 2017.

WU, D. *et al.* Iterative low-dose CT reconstruction with priors trained by artificial neural network. **IEEE Transactions on Medical Imaging**, IEEE, v. 36, n. 12, p. 2479–2486, 2017.

WU, G.; MAINPRIZE, J. G.; YAFFE, M. J. Dose reduction for digital breast tomosynthesis by patch-based denoising in reconstruction. *In*: SPRINGER. **International Workshop on Digital Mammography**. [*S.l.: s.n.*], 2012. p. 721–728.

WU, G.; MAINPRIZE, J. G.; YAFFE, M. J. Spectral analysis of mammographic images using a multitaper method. **Medical Physics**, Wiley Online Library, v. 39, n. 2, p. 801–810, 2012.

WU, W. *et al.* DRONE: Dual-domain residual-based optimization Network for sparse-view CT reconstruction. **IEEE Transactions on Medical Imaging**, IEEE, 2021.

WÜRFL, T. *et al.* Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1454–1463, 2018.

XIA, W. *et al.* Physics-/model-based and data-driven methods for low-dose computed tomography: A survey. **IEEE Signal Processing Magazine**, IEEE, v. 40, n. 2, p. 89–100, 2023.

XU, S. *et al.* Statistical iterative reconstruction to improve image quality for digital breast tomosynthesis. **Medical Physics**, Wiley Online Library, v. 42, n. 9, p. 5377–5390, 2015.

YAFFE, M. J. Digital mammography. *In*: METTER, R. L. V.; BEUTEL, J.; KUNDEL, H. L. (ed.). **Handbook of Medical Imaging: Volume 1. Physics and Psychophysics**. 1. ed. Bellingham: SPIE Press, 2000. cap. 5, p. 331–354.

YAFFE, M. J.; MAINPRIZE, J. G. Risk of radiation-induced breast cancer from mammographic screening. **Radiology**, Radiological Society of North America, Inc., v. 258, n. 1, p. 98–105, 2011.

YANG, Q. *et al.* Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1348–1357, 2018.

YANG, X. *et al.* Low-dose x-ray tomography through a deep convolutional neural network. **Scientific Reports**, Nature Publishing Group, v. 8, n. 1, p. 1–13, 2018.

YIN, X. *et al.* Domain progressive 3D residual convolution network to improve low-dose CT imaging. **IEEE Transactions on Medical Imaging**, IEEE, v. 38, n. 12, p. 2903–2913, 2019.

YOU, C. *et al.* CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). **IEEE Transactions on Medical Imaging**, IEEE, v. 39, n. 1, p. 188–203, 2019.

ZHANG, M. *et al.* VST-net: variance-stabilizing transformation inspired network for poisson denoising. **Journal of Visual Communication and Image Representation**, Elsevier, v. 62, p. 12–22, 2019.

ZHANG, Q.; WU, Y. N.; ZHU, S.-C. Interpretable convolutional neural networks. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2018. p. 8827–8836.

ZHANG, Y.; YU, H. Convolutional neural network based metal artifact reduction in x-ray computed tomography. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 6, p. 1370–1381, 2018.

ZHAO, H. *et al.* Loss functions for image restoration with neural networks. **IEEE Transactions on Computational Imaging**, IEEE, v. 3, n. 1, p. 47–57, 2016.

ZHENG, J.; FESSLER, J. A.; CHAN, H.-P. Detector blur and correlated noise modeling for digital breast tomosynthesis reconstruction. **IEEE Transactions on Medical Imaging**, IEEE, v. 37, n. 1, p. 116–127, 2017.

ZHU, J.-Y. *et al.* Unpaired image-to-image translation using cycle-consistent adversarial networks. *In*: **Proceedings of the IEEE International Conference on Computer Vision**. [*S.l.: s.n.*], 2017. p. 2223–2232.

EESC • USP