Deep Convolutional Neural Network With Adversarial Training for Denoising Digital Breast Tomosynthesis Images

Mingjie Gao, *Graduate Student Member, IEEE*, Jeffrey A. Fessler, *Fellow, IEEE*, and Heang-Ping Chan

***Abstract*— Digital breast tomosynthesis (DBT) is a quasi-three-dimensional imaging modality that can reduce false negatives and false positives in mass lesion detection caused by overlapping breast tissue in conventional two-dimensional (2D) mammography. The patient dose of a DBT scan is similar to that of a single 2D mammogram, while acquisition of each projection view adds detector readout noise. The noise is propagated to the reconstructed DBT volume, possibly obscuring subtle signs of breast cancer such as microcalcifications (MCs). This study developed a deep convolutional neural network (DCNN) framework for denoising DBT images with a focus on improving the conspicuity of MCs as well as preserving the ill-defined margins of spiculated masses and normal tissue textures. We trained the DCNN using a weighted combination of mean squared error (MSE) loss and adversarial loss. We configured a dedicated x-ray imaging simulator in combination with digital breast phantoms to generate realistic *in silico* DBT data for training. We compared the DCNN training between using digital phantoms and using real physical phantoms. The proposed denoising method improved the contrast-to-noise ratio (CNR) and detectability index (d') of the simulated MCs in the validation phantom DBTs. These performance measures improved with increasing training target dose and training sample size. Promising denoising results were observed on the transferability of the digital-phantom-trained denoiser to DBT reconstructed with different techniques and on a small independent test set of human subject DBT images.**

***Index Terms*— Deep convolutional neural network, digital breast tomosynthesis, generative adversarial network, image denoising, microcalcification.**

I. Introduction

DIGITAL breast tomosynthesis (DBT) is an important imaging modality for breast cancer screening and diagnosis. A DBT system acquires a sequence of projection views (PVs) within a limited angle [1]–[4]. A quasi-three-dimensional (3D) volume is reconstructed from the two-dimensional (2D) PVs to reduce the superimposition of breast tissues that can cause false negatives and false positives in 2D mammography. The patient dose of a DBT scan is similar to that of a single 2D mammogram, while acquisition of each PV adds detector readout noise. The noise is propagated to the DBT volume through reconstruction, which may obscure subtle signs of breast cancer.

In breast imaging, the important signs of breast cancer manifest as mass, architectural distortion, and clustered microcalcifications (MCs). Malignant masses are low-contrast objects with ill-defined margins or irregular shapes. Clinically significant MCs seen on breast x-ray images have diameters of less than about 0.5 mm. Although MCs contain calcium that has relatively high x-ray attenuation, the small sizes can result in overall low conspicuity. It is a challenge to denoise DBT images because conventional noise smoothing methods may also smooth out the subtle MCs. Our focus is to improve the conspicuity of MCs and preserve the natural appearance of soft tissues and masses in DBT images.

Researchers have tried various methods to suppress noise in DBT images. PV filtration was performed using a linear filter [5] or a neural network filter with one convolutional layer [6]. Model-based iterative reconstruction (MBIR) has attracted much attention because of its potential of handling noise. Statistical noise models, such as noise variance [7]–[9], detector blur and correlated noise (DBCN) [10] and scattered noise [11], were incorporated into the DBT system model for MBIR. Gradient-based regularizers, such as selective-diffusion regularizer [12], total variation (TV) [13], [14] and its variants [15]–[17], were also used in MBIR of DBT for noise reduction. However, MBIR techniques may introduce "plastic appearance" to the soft tissue structures as observed in CT reconstruction [18], [19]. Several denoising methods were proposed for the reconstructed DBT images. Das *et al*. used a 3D Butterworth filter to improve MC detection [20]. Abdurahman *et al*. iteratively applied a smoothing filter to improve the contrast-to-noise ratio (CNR) of MCs [21]. Lu *et al*. applied multiscale bilateral filtering [22] either to the reconstructed images as post-processing or between reconstruction iterations, improving the CNRs of MCs without distorting the masses.

Manuscript received February 18, 2021; accepted March 14, 2021. Date of publication March 17, 2021; date of current version June 30, 2021. This work was supported by the National Institutes of Health under Award R01 CA214981. *(Corresponding author: Mingjie Gao.)*

Mingjie Gao and Jeffrey A. Fessler are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA, and also with the Department of Radiology, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: gmingjie@umich.edu; fessler@umich.edu).

Heang-Ping Chan is with the Department of Radiology, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: chanhp@umich.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3066896, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3066896

Recently, deep convolutional neural network (DCNN) methods have shown state-of-the-art performances in natural image restoration tasks. Zhang *et al*. constructed a feed-forward denoising convolutional neural network (DnCNN) for Gaussian noise removal [23]. The DCNN training loss was the mean squared error (MSE) between the network output and clean training target. Dong *et al*. trained a three-layer convolutional network [24] and Kim *et al*. trained a 20-layer network [25] with MSE loss for single-image super-resolution. However, there is a perception-distortion tradeoff: the MSE loss tends to produce overly smoothed images that are not visually satisfactory even if their MSE, peak signal-to-noise ratio or structural similarity are high [26]. Alternative training losses were designed to address this problem. For example, Johnson *et al*. used feature-level MSE loss, called the perceptual loss, for image transformation and super-resolution [27]. Inspired by the generative adversarial network (GAN) [28], Ledig *et al*. introduced the adversarial loss for image super-resolution and greatly increased the mean opinion scores [29]. The adversarial training stability was further improved as the Wasserstein GAN (WGAN) was proposed [30], [31]. The adversarial loss was applied to medical image processing and achieved promising results, including CT denoising and artifacts correction [32]–[36] and MRI dealiasing [37], [38]. We previously used a DCNN to denoise PVs before DBT reconstruction and achieved moderate CNR improvement for MCs [39]. In this study, we trained a DCNN using a weighted combination of MSE loss and adversarial loss to denoise reconstructed DBT images.

A DCNN having millions of parameters requires a large amount of data to learn complex image patterns. However, in medical imaging fields, training data is limited due to the high costs of collecting and annotating the data. For a denoising task using a supervised approach, the DCNN training requires high dose (HD) images as references or targets to learn to reduce noise of a corresponding input low dose (LD) images, but we cannot scan a patient with a HD technique. To overcome these problems, we studied the feasibility of using two methods for generating data to train DCNN for DBT denoising. The first method is to generate *in silico* training data. The virtual imaging clinical trial for regulatory evaluation (VICTRE) project [40] conducted a computer-simulated imaging trial to evaluate DBT as a replacement for digital mammography. It provides an anthropomorphic breast model to generate digital breast phantoms[1] [41]. We incorporated the digital breast phantom into an x-ray imaging simulation tool, developed by GE Global Research, named Computer Assisted Tomography SIMulator[2] (CatSim) [42], [43], to generate relatively realistic breast images from a clinical DBT system and use them for DCNN training. The second method is to prepare physical heterogeneous breast phantoms using tissue-mimicking materials and scan them with a DBT imaging system [3], [4]. We trained the DCNNs using the two types of data and compared their denoising performances.

[1]Available at https://github.com/DIDSR/VICTRE.

[2]We used an earlier version of CatSim that is Matlab-based. A Python-based CatSim is recently available at https://github.com/xcist/CatSim.

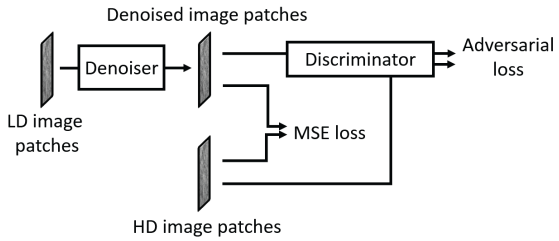


Fig. 1. The framework of the denoising DCNN with adversarial training (DNGAN).

The paper is organized as follows. Section II introduces our DCNN training framework, the data sets, and the figures of merit. Section III investigates the effects of training set properties and the hyper-parameters and presents the denoising results. Section IV discusses the advantages of our denoising approach and the limitations of our study. Section V concludes the paper.

II. Methods and Materials

A. DCNN Training

1) DNGAN Framework: To reduce the noise in the reconstructed DBT images, we would like to obtain a mapping function, called a denoiser, that maps the noisy images to clean or less noisy ones. The denoiser was implemented as a DCNN with trainable weights. In the training phase, the denoiser learned how to denoise by adjusting the trainable weights to minimize the training loss function. In the deployment phase, the denoiser with frozen weights served as a well-trained function that could be applied to noisy DBT images.

During training, the system took pairs of LD and HD images as input and target, respectively. The HD target images were used to guide the denoiser to generate denoised images from the LD images by minimizing a weighted combination of MSE loss and adversarial loss, where the adversarial loss was derived by training a discriminator to distinguish between the denoised LD and the target HD images as in a GAN. As demonstrated below, the GAN-based adversarial training was crucial to constrain the degree of smoothing and maintain the sharpness of the denoised DBT images. We therefore call our training framework DNGAN. We chose to use LD/HD image regions, or patches, 32×32 pixels in size as the DNGAN inputs to allow the DCNN to focus on the local image structures in the adversarial training [44]. Fig. 1 shows the framework of the DNGAN. Section S-I of the Supplementary Materials describes the network structures of the denoiser and the discriminator.

2) Training Loss Function: The training loss function is composed of the MSE loss L_{MSE} and the adversarial loss L_{adv}

$$\underset{G}{\operatorname{argmin}}\, L_{\text{MSE}}(G) + \lambda_{\text{adv}} \cdot L_{\text{adv}}(G) \tag{1}$$

where G denotes the denoiser, λ_{adv} is a tuning parameter controlling the weighting between the MSE loss, which contributes to image smoothness, and the adversarial loss, which contributes to preserving high frequency image textures.

The MSE loss compares the pixel-wise difference between the denoised image patches and the corresponding HD target image patches x_{target} as follows

$$L_{\text{MSE}}(G) = \mathbb{E}_{(x_{\text{noisy}}, x_{\text{target}})}\left[\frac{1}{N_{\text{pixel}}}\left\|G\left(x_{\text{noisy}}\right) - x_{\text{target}}\right\|^2\right] \quad (2)$$

where x_{noisy} is the LD noisy input patch, N_{pixel} is the number of pixels in an image patch.

We implemented the adversarial loss as the WGAN with gradient penalty [31]. Section S-II of the Supplementary Materials summarizes the key idea behind the derivation of the adversarial loss. The Wasserstein distance (WD) is estimated as suggested in [30]

$$\hat{d}_G(D) := \mathbb{E}_{x_{\text{target}}}\left[D\left(x_{\text{target}}\right)\right] - \mathbb{E}_{x_{\text{noisy}}}\left[D\left(G\left(x_{\text{noisy}}\right)\right)\right] \quad (3)$$

where D is the discriminator, or a critic, whose output is a similarity score that assesses whether the input image patch comes from the distribution of the target images. The training loss function for the discriminator is

$$\underset{D}{\text{argmin}} -\hat{d}_G(D) + \lambda_D \cdot \mathbb{E}_{\bar{x}}\left[\left(\|\nabla_{\bar{x}} D(\bar{x})\| - 1\right)^2\right] \quad (4)$$

where $\bar{x} = t \cdot x_{\text{target}} + (1-t) \cdot G(x_{\text{noisy}})$ is an interpolated image, $t \sim \text{Unif}([0,1])$, λ_D is the penalty weight. To promote the denoised images to be perceptually similar to the target images, we maximize the term $\mathbb{E}_{x_{\text{noisy}}}\left[D\left(G\left(x_{\text{noisy}}\right)\right)\right]$ in $\hat{d}_G(D)$ when optimizing G, which gives the corresponding adversarial loss of the denoiser training

$$L_{\text{adv}}(G) = -\mathbb{E}_{x_{\text{noisy}}}\left[D\left(G\left(x_{\text{noisy}}\right)\right)\right]. \quad (5)$$

3) Fine-Tuning With MC Patches: For DNGAN training, the DBT images with any kind of breast structures can be used for training, as long as the LD/HD image pairs contain matched structures to be preserved and residual differences to be reduced. In our training set preparation, we extracted non-overlapping patches from the DBT slices using a shifting window. A DBT volume mainly consisted of tissue background and there were very few MCs in a volume. Consequently, the training set was dominated by background patches. To emphasize the MC images so that the denoiser could focus on the MC signals and learn to preserve or enhance them, we investigated the feasibility of a second training stage that fine-tuned the DNGAN only with patches centered at individual MCs.

In the fine-tuning stage, we adopted the training technique of layer freezing [45]. For example, freezing m layers means that layers 1 to m in the denoiser were frozen, and layers $(m+1)$ to the last layer were active. Our denoiser had a total of 10 layers, so $m \leq 10$. Note that the freezing was only applied to the denoiser network. All layers in the discriminator were active during training.

B. Data Sets

We prepared several data sets to investigate the effects of the dose level of the HD training target, the training sample size, and the underlying reconstruction algorithm on the performance of our proposed DNGAN. We also prepared a data set for the aforementioned MC fine-tuning experiment. Table I summarizes the data sets and their use in the experiments.

As introduced in Section I, we generated two types of data, namely the digital phantom data and the physical phantom data, for DNGAN training and validation. The digital phantom data provided a wide range of x-ray exposures including noiseless images for the study of the effect of the dose level of the HD training target. We also used the digital data to generate the MC fine-tuning set because the coordinates of the simulated MCs were known exactly. Compared with the digital phantom data, the physical phantom data contained all the imaging degradation factors of a DBT system and were considered to be more realistic, and only physical phantom data imaged with the DBT system could be reconstructed with the manufacturer's proprietary reconstruction technique. We therefore used the physical phantom data for the other experiments and for study of the transferability of a digital-data-trained DNGAN denoiser to real DBT images acquired from physical phantoms and human subjects.

1) Digital Phantom Data: We prepared 25 heterogeneous dense (34% glandular volume fraction) 4.5-cm-thick digital phantoms at a voxel resolution of 0.05 mm using the VICTRE breast model [40], [41]. We inserted simulated MCs consisting of calcium oxalate in clusters into the digital phantoms. Each cluster had 12 MCs arranged on a 3-by-4 grid parallel to the detector plane with a small offset in the direction perpendicular to the chest wall to avoid in-plane artifacts interfering each other during reconstruction. The MCs had three diameters: 0.150 mm, 0.200 mm, 0.250 mm. We used 24 phantoms for training data preparation and held out one phantom for validation.

Next we configured CatSim [42], [43] to model the Pristina DBT system (GE Healthcare) that acquires 9 PVs in 25° scan. Section S-III of the Supplementary Materials provides the detailed description of the CatSim configuration. To simulate LD PVs for the digital phantoms, we set the total x-ray exposure of 9 PVs to 24 mAs in CatSim, which was close to the value from automatic exposure control (AEC) for a 4.5 cm breast for the Pristina system [46]. We reconstructed the DBT volumes at a voxel size of 0.1 mm × 0.1 mm × 1 mm using three iterations of simultaneous algebraic reconstruction technique (SART) [7] with the segmented separable footprint projector [47].

2) Physical Phantom Data: We used seven 1-cm-thick heterogeneous slabs with 50% glandular/50% adipose breast-tissue-equivalent material to construct the physical phantoms [3], [4]. By arranging five slabs in different orders and orientations, we formed nine 5-cm-thick phantoms. Clusters of simulated MCs (glass beads) of three nominal diameters (0.150-0.180 mm, 0.180-0.212 mm, 0.212-0.250 mm) were randomly sandwiched between the slabs. Glass beads are used in some commercial breast phantoms but have lower x-ray attenuation than calcium oxalate specks of the same size. We used eight phantoms for training data preparation and held out one phantom for validation.

Each phantom was scanned twice, one at LD and the other at HD, by a Pristina DBT system under the same compression. The LD scans were acquired with the standard

TABLE I
A SUMMARY OF THE DATA SETS

Purpose	Name	Phantom type	Recon algorithm	No. of patch pairs	Comments
Training	24mAs/target	Digital	SART	199,850	target = 72mAs, 120mAs, 360mAs, noiseless. For investigating the effect of the dose level of the HD training target.
	MC fine-tuning set	Digital	SART	3,048	Patches centered at individual MCs generated at known locations. For investigating the feasibility of a second fine-tuning stage.
	LD/HD-k	Physical	SART	$k \times 400{,}000$	k = 20%, 35%, 50%, 65%, 80%, 100%. For investigating the effect of the training sample size.
	LD/HD-Pristina	Physical	Pristina	400,000	For training a matched denoiser when evaluating the generalizability of DNGAN in terms of the reconstruction algorithms.
Validation	24mAs as input, higher dose levels as reference truth	Digital	SART	/	Has ground truth scans simulated at multiple dose levels. Used for NPS comparison.
	LD as input, HD as reference for performance comparison	Physical	SART	/	Has individually marked MCs of three nominal diameters. Used for CNR, FWHM, fit success rate, d', and visual comparisons.
			Pristina	/	For evaluating the generalizability of DNGAN in terms of the reconstruction algorithms.
Test	Human subject DBTs	/	SART	/	An independent test set. For demonstrating the robustness and the feasibility of applying a denoiser trained with phantom data to human DBTs.

dose (STD) setting, which automatically chose a technique of Rh/Ag 34 kVp. The exposures ranged between 30.4 mAs and 32.6 mAs with a mean of 31.4 mAs for the nine phantoms. We manually set the exposure for the HD scans to Rh/Ag 34 kVp, 125 mAs. The reconstruction parameters were the same as those for the digital phantoms.

We marked the MCs in the SART-reconstructed HD volume of the hold-out validation phantom for denoiser evaluation. There was a total of 236 MCs of size 0.150-0.180 mm, 227 MCs of 0.180-0.212 mm, and 159 MCs of 0.212-0.250 mm.

3) Training Set Generation:

a) Training sets with different target dose levels: Using CatSim with the digital phantoms, we prepared HD images over a range of dose levels to study the effect of the dose level of the training target images on the effectiveness of the trained denoiser. Specifically, we simulated the HD DBT scans with 72 mAs (3× AEC), 120 mAs (5× AEC), 360 mAs (15× AEC) and noiseless (∞× AEC) settings. We paired these HD scans with the 24 mAs scans to form four training sets, referred to as 24mAs/72mAs, 24mAs/120mAs, 24mAs/360mAs and 24mAs/noiseless, respectively. We extracted 199,850 pairs of patches from the 24 pairs of SART-reconstructed DBT volumes of the digital phantoms as the training set for each dose condition.

b) MC Fine-Tuning Set: We prepared the training set with patches centered at each MC for fine-tuning using the digital phantom data. There were 1,032 MCs of 0.150 mm, 1,008 MCs of 0.200 mm, 1,008 MCs of 0.250 mm in the 24 digital phantoms, giving a total of 3,048 MC patches in the fine-tuning set.

c) Training Sets With Different Sample Sizes: The generalizability of a trained DCNN depends on the training sample size [48]. We designed an experiment to study the effect of training sample sizes for the DBT denoising task using the physical phantom data. Specifically, we first extracted 400,000 pairs of patches from the eight physical phantoms to form the pool of training patches. Then we randomly drew 20%, 35%, 50%, 65%, 80% of patches from the pool to simulate five training set sizes in addition to the 100% set. These training sets were referred to as LD/HD-k, where k is the drawing percentage. The subset drawing at each percentage was repeated 10 times with different random seeds. Although the independence among the drawn subsets decreased as k increased, the simulation study would provide some understanding of the trend and variation of the training.

d) Training Set of Pristina-Reconstructed Images: The Pristina DBT system has a built-in commercial reconstruction algorithm. We refer to it as Pristina algorithm. To evaluate the generalizability of the DNGAN denoiser in terms of the reconstruction algorithms, we directly deployed the denoiser that was trained with SART-reconstructed images to the Pristina-reconstructed images. We also prepared a training set using Pristina-reconstructed images to train a matched denoiser for comparison. This training set was referred to as LD/HD-Pristina and had 400,000 pairs of patches extracted from the eight training physical phantoms.

Section S-IV of the Supplementary Materials shows two examples of the training patches. For all the training sets, we subtracted the mean from each DBT volume to center its histogram before patch extraction.

4) Human Subject DBTs:

We used eleven de-identified human subject DBT scans, previously collected for another study with IRB approval, as an independent test set to evaluate the denoising effect on the CNR of MCs and the appearances of breast tissue and cancerous masses in real breasts. They contained biopsy-proven invasive ductal carcinomas (masses) and ductal carcinomas in situ (MC clusters). The images were acquired using a GE prototype GEN2 DBT system. The prototype system acquired 21 PVs in a scan angle of 60°.

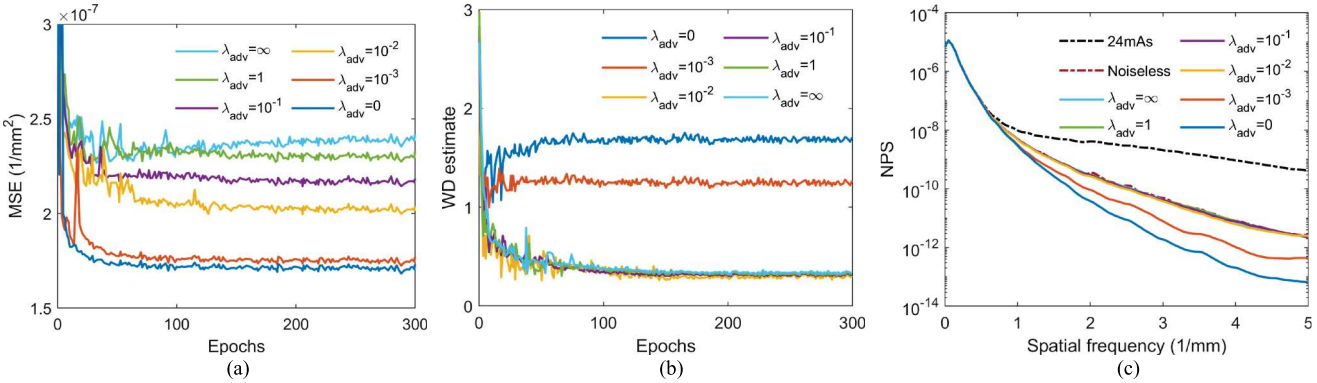


Fig. 2. (a) The training MSE losses and (b) the WD estimates for different λ_{adv} versus training epochs. (c) The NPS curves of the denoised validation digital phantom volumes for different λ_{adv}.

We used the central 9 PVs that corresponded to a scan angle of 24° to simulate an LD DBT with scan parameters like the Pristina DBT system. The DBTs were reconstructed with SART with the same parameters as described above. 301 MCs were marked in the DBT volumes.

C. Figures of Merit

The structural noise power spectrum (NPS) of a breast image quantifies both the structured noise of the object being imaged and other noise on the imaging chain. It has been shown to have a power-law form for mammograms [49]. We used the NPS to quantify the change of textures and noise in the denoised DBT.

To quantitatively evaluate the MCs in the images, we calculated the contrast-to-noise ratio (CNR) and full width at half maximum (FWHM) by fitting a Gaussian function to each MC. For comparison of the performance of the denoiser at different conditions, we calculated the mean and standard deviation of the CNR and the FWHM over the marked MCs at each speck size on the validation physical phantom. The denoiser inevitably smoothed out some subtle MCs as if they were noise. If an MC was very blurred, the fitting program would fit to the background. We set three criteria to automatically mark the fitting as a failure: if the FWHM was larger than twice of the nominal MC size, or if the fitted Gaussian was off centered by two pixels, or if the fitting error was larger than a threshold. These failed MCs were excluded from the mean CNR or mean FWHM calculations, but the fit success rate was counted for each MC speck size and considered one of the indicators of the denoiser performance.

We also calculated the task-based detectability index (d') from the nonprewhitening matched filter model observer with eye filter (NPWE) as an image quality metric [50], [51]. The NPWE observer performance was shown to correlate well with human observer performance [51]–[55]. In this study, we considered the task of detecting MCs of different nominal sizes in the heterogeneous background of breast phantom DBT.

Section S-V of the Supplementary Materials describes the calculation details of the NPS, CNR, FWHM, and d'.

D. DCNN Training Setup

For the DNGAN training, we randomly initialized all the kernel weights. We set the mini-batch size to 512, and λ_D to 10 as suggested [31]. We set $\lambda_{\text{adv}} = 10^{-2}$. The denoiser and the discriminator were trained alternately, and the discriminator had 3 steps of updates for every step of the denoiser update. The discriminator and the denoiser both used Adam optimizer [56] and shared the same learning rate. The learning rate started with 10^{-3} and dropped by a factor of 0.8 for every 10 epochs. The learning rate started with 10^{-4} and dropped by a factor of 0.8 for every 50 epochs in the fine-tuning stage. We selected 300 epochs for stage one training and 1,000 epochs for fine-tuning. The selection of λ_{adv} is shown in Section III.A. The other parameters, batch size, learning rate, and the number of epochs, were also chosen experimentally based on the training convergence and efficiency as shown in Section S-VI of the Supplementary Materials. The DCNN model was implemented in Python 2.7 and TensorFlow 1.4.1. The training was run on one Nvidia GTX 1080 Ti GPU.

E. Comparison Method

We included a DBT MBIR algorithm developed in our laboratory that models the detector blur and correlated noise (DBCN) with an edge-preserving regularizer [10] for comparison. The parameters ($\beta = 70$, $\delta = 0.002$/mm, 10 iterations) that were chosen in [10] were used for reconstructing the DBTs from the GE prototype system. We adapted the DBCN to the Pristina system by adjusting β to 40.

III. Results

A. Effect of Tuning Parameter λ_{adv}

To demonstrate the effect of λ_{adv} in (1), we trained six denoisers using $\lambda_{\text{adv}} = 0$, 10^{-3}, 10^{-2}, 10^{-1}, 1, ∞ in the DNGAN. The condition $\lambda_{\text{adv}} = 0$ is equivalent to using the MSE loss only, and the condition $\lambda_{\text{adv}} = \infty$ is equivalent to using the adversarial loss only for DNGAN training. The training set was 24mAs/noiseless. We used the same random seeds for weight initialization and data batching for all conditions.

Fig. 2(a) and (b) shows the training MSE losses and the WD estimate $\hat{d}_G(D)$ defined in (3) versus training epochs. For small λ_{adv} (0 and 10^{-3}), even though they converged to low training MSE values, they had high WD estimates which means that the denoisers produced images that were perceptually dissimilar to the noiseless targets in the training. Note that the WD estimate $\hat{d}_G(D)$ could be increasing or decreasing versus training epochs because the adversarial training aimed at maximizing it over D and minimizing it over G. Fig. 2(c) shows that the denoised validation digital

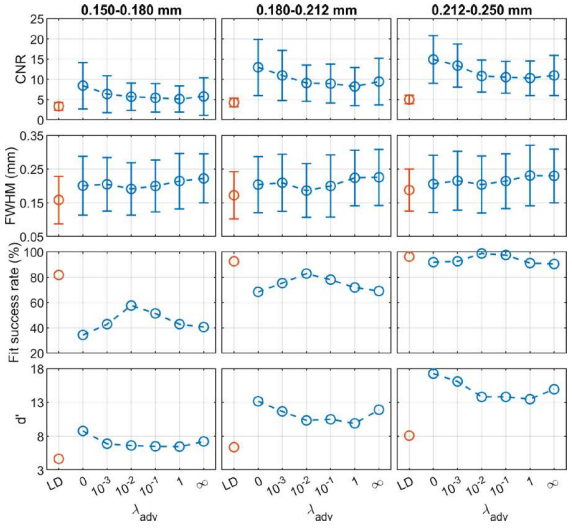


Fig. 3. The CNR, FWHM, fit success rates, and d' of the MCs in the validation physical phantom for different λ_{adv}. The error bars represent one standard deviation.

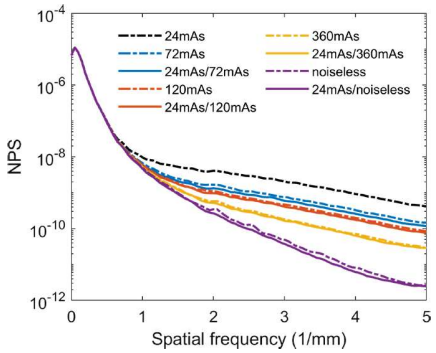


Fig. 4. The NPS curves of the validation digital phantom volumes for comparing the dose levels of the training targets. The solid lines are calculated from the volumes by deploying the denoisers to the 24mAs volume. The dashed lines are from the CatSim simulated volumes with the corresponding dose levels.

phantoms of $\lambda_{adv} = 0$ and 10^{-3} had low NPS compared to the ground truth noiseless image. This is evidence that the images were overly smoothed and lost structural details.

For $\lambda_{adv} = 10^{-2}$, 10^{-1}, 1, ∞, although these conditions had similar WD estimates and NPS, the converged training MSE values monotonically increased as λ_{adv} increased in Fig. 2(a). Moreover, as shown in Fig. 3, for the MCs in the denoised validation physical phantoms, the FWHM increased and the fit success rate dropped substantially as λ_{adv} increased beyond 10^{-2}, indicating the blurring and loss of MC signals. Therefore, when the image smoothness was comparable, we preferred $\lambda_{adv} = 10^{-2}$ for a smaller training MSE and MC preservation.

B. Effect of Dose Level of Targets on DNGAN Training

To study the effect of the dose level of the training target images on the effectiveness of the denoiser, we trained the DNGAN using the 24mAs/72mAs, 24mAs/120mAs, 24mAs/360mAs, 24mAs/noiseless sets with $\lambda_{adv} = 10^{-2}$. We used the same random seeds for weight initialization and data batching for all conditions.

Fig. 4 shows that the NPS of the denoised validation digital phantoms matched the NPS of the corresponding CatSim-simulated ground truth volumes. Fig. 5 shows that the

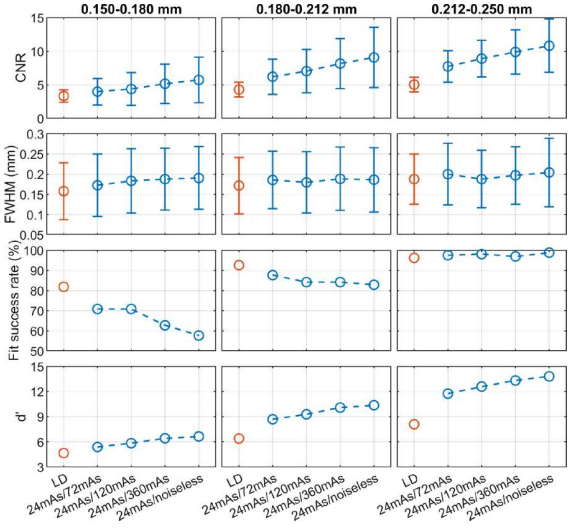


Fig. 5. The CNR, FWHM, fit success rates, and d' of the MCs in the validation physical phantom for the different dose levels of the targets used in the DNGAN training.

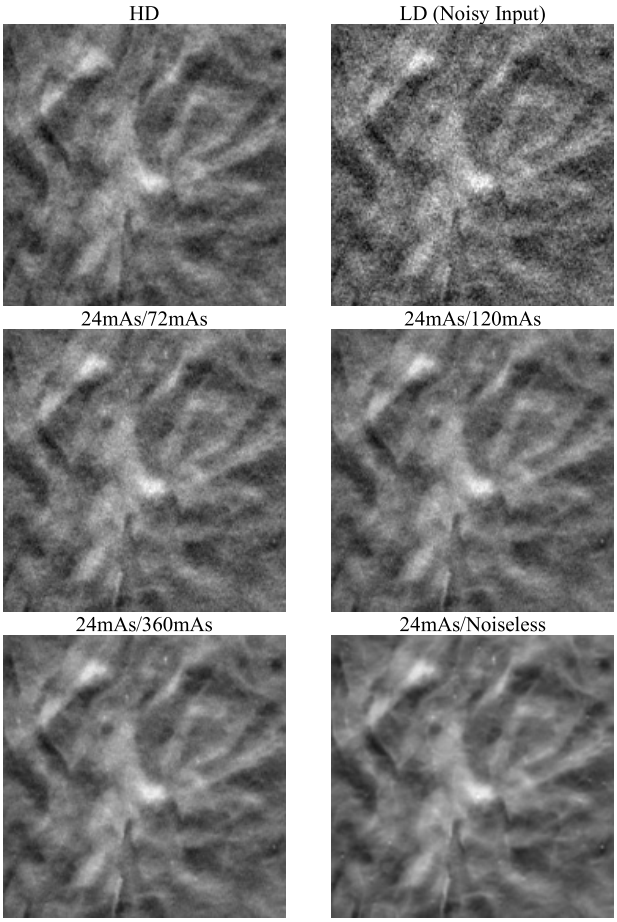


Fig. 6. An example 18 mm × 18 mm region in the validation physical phantom for the different dose levels of the targets used in the DNGAN training. The images are displayed with the same window/level settings. The HD scan of the validation physical phantom (34 kVp, 125 mAs) is also shown for reference.

DNGAN achieved higher CNR and d' for the MCs in the validation physical phantoms when the training targets were acquired with a higher dose. However, as the target dose level increased, the fit success rate decreased for the two smaller MC groups. Fig. 6 shows that if the target dose level was very high, for example, 360 mAs or infinity, the tissue backgrounds

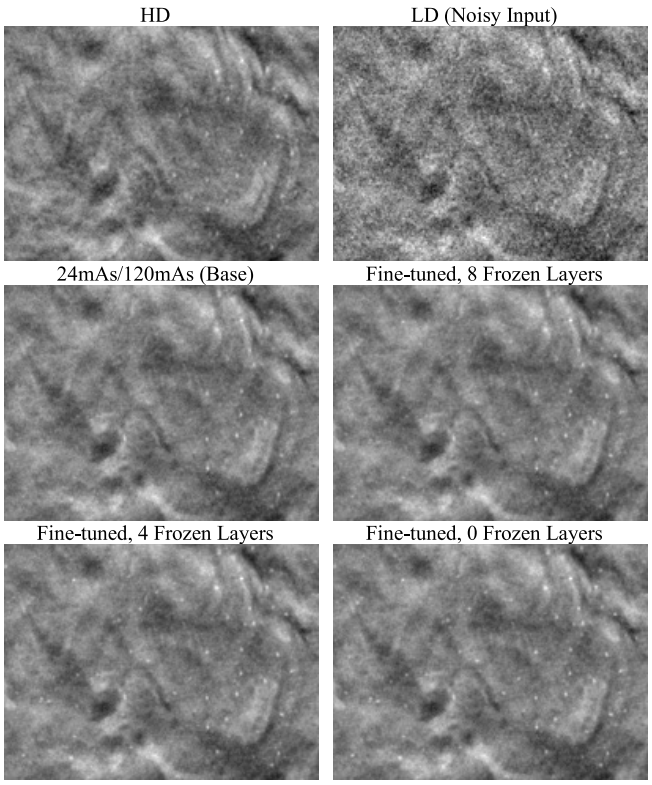


Fig. 7. An example 20 mm × 15 mm region in the validation physical phantom showing the effect of fine-tuning and layer freezing. The region contains a background area and a 0.180-0.212 mm MC cluster. The images are displayed with the same window/level settings.

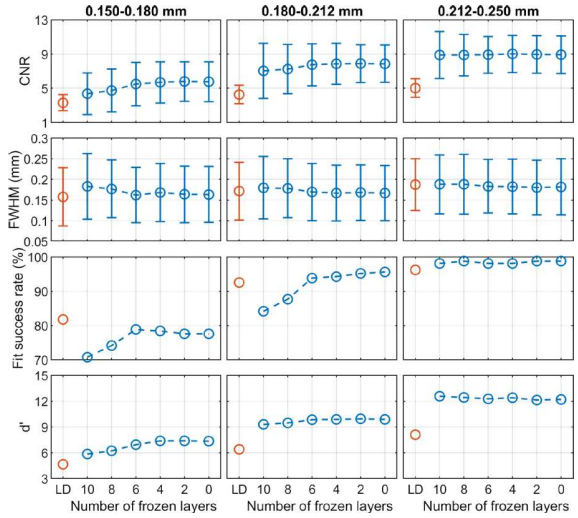


Fig. 8. The CNR, FWHM, fit success rates, and d' of the MCs in the validation physical phantom showing the effect of fine-tuning and layer freezing.

of the denoised validation physical phantoms images could be too smooth. The smoothing effect of the 24mAs/noiseless denoiser is further demonstrated in human subject DBTs in Section III.F. We used the 24mAs/120mAs for training the DNGAN in the following studies since its dose ratio was closer to that of the LD/HD physical phantom images from a real scan.

C. Effect of MC Fine-Tuning and Layer Freezing

We selected the DNGAN trained with 24mAs/120mAs in Section III.B as the initial model and fine-tuned it using the

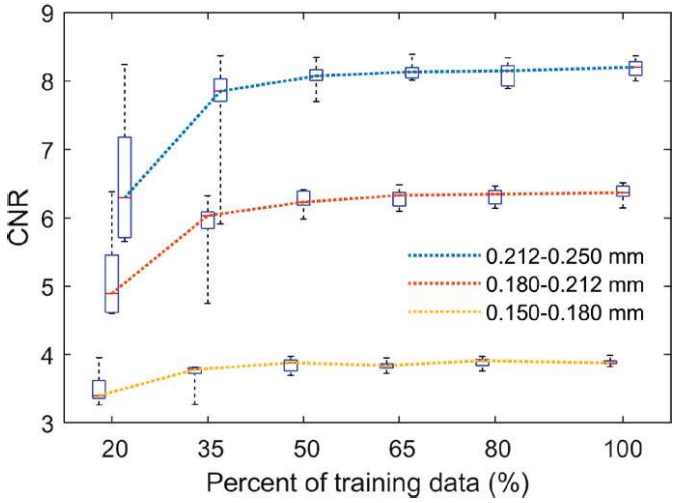


Fig. 9. The box plots of the CNRs for different training sample sizes. Each box contains 10 data points. In the box plot, the red bar represents the median; the length of the box equals the interquartile range; the whiskers extend to the minimum and maximum data points. The boxes are slightly shifted horizontally to avoid overlap.

MC data set. We set the number of frozen layers of the denoiser to 8, 6, 4, 2, 0 while all layers in the discriminator were allowed to be fine-tuned under all conditions. We obtained five fine-tuned denoisers in addition to the base denoiser that was equivalent to freezing 10 layers.

Fig. 7 shows that the fine-tuned denoisers improved the visibility of the subtle MCs in the denoised validation physical phantoms compared to the base denoiser. Fig. 8 shows that the CNR and d' values increased and the FWHM values decreased as the number of frozen layers decreased, indicating that the MCs became brighter and sharper. The fit success rate also increased for the two smaller MC groups. However, the improvements leveled off when fewer than about 6 layers were frozen. In addition, as seen in the examples in Fig. 7, the fine-tuning not only enhanced the subtle MCs but also some MC-like noise and background structures in the denoised images. The false MCs were obvious and distracting for all the fine-tuned denoisers even though freezing layers mitigated the problem to some extent. The fine-tuning was excluded from further discussions below because we concluded that it was unsuitable for practical use at this point.

D. Effect of Training Sample Sizes

We trained the DNGAN using the LD/HD-*k* datasets from the physical phantoms. A different random seed was used for the weight initialization and data batching in each repeated experiment to account for the training randomness. After training, we deployed the denoiser to the validation physical phantom and calculated the mean CNRs for the MCs.

Fig. 9 shows box plots using the 10 repeated experiments versus training data percentage. The general trend was that, when the training sample size increased, the training variation became smaller, and the median CNR increased and became stable. The CNR variations were large at 20% and 35%. This is especially undesirable for DBT because a denoiser with large performance variations can have unpredictable effect on subtle MCs. The large variation can be attributed mainly to

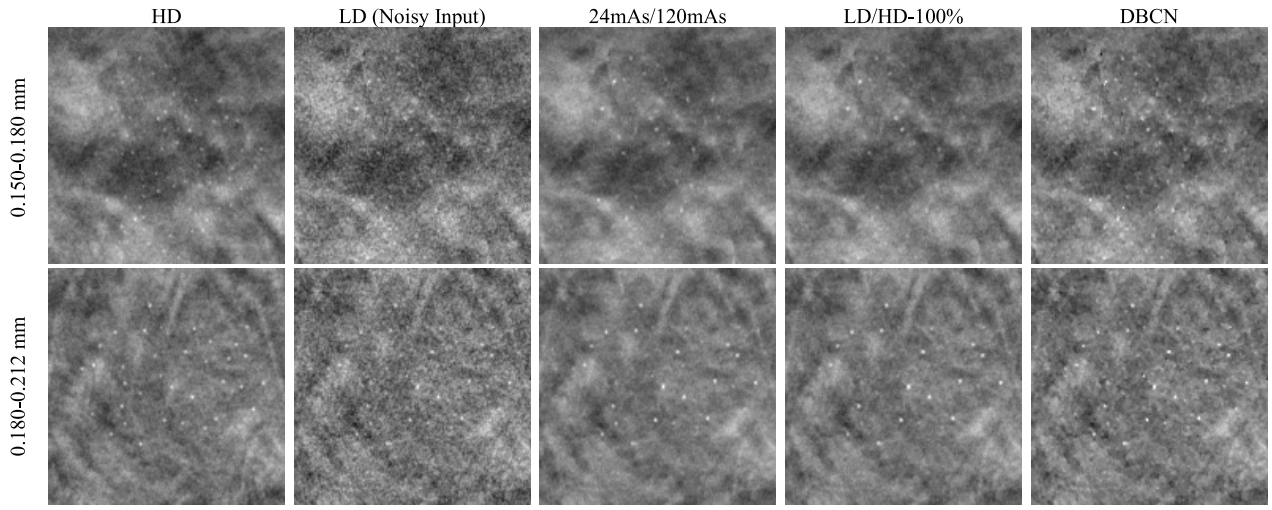


Fig. 10. Example MC clusters in the validation physical phantom for comparing the denoising results. All images show a 15 mm × 15 mm region. The images in the same row are displayed with the same window/level settings.

the insufficient representation of the imaging characteristics by the small training set and the overfitting of the DCNN to each set of samples. The training randomness from weight initialization and data batching also contributed substantially to the variations, as can be seen from the 100% data point where the training set was the same for all repeated experiments.

E. Denoising Performance on Validation Physical Phantom

We compared the DNGAN-denoised LD images where the DNGAN was trained with the digital phantom data (24mAs/120mAs) or the physical phantom data (LD/HD-100%), and the LD images reconstructed from the DBCN algorithm. The LD/HD-100% model that was closest to the mean performance among the 10 repeated trainings in Section III.D was used in this comparison.

Fig. 10 shows that the backgrounds in the 24mAs/120mAs and LD/HD-100% denoised validation physical phantoms were perceptually similar to the HD references, with the former being less noisy. Both denoisers improved the CNRs significantly ($p < 0.001$ for all three MC sizes, two-tailed paired t-test) compared to the LD images, as shown in Fig. 11. Moreover, 24mAs/120mAs had significantly higher CNRs than LD/HD-100% ($p < 0.001$ for all three MC sizes) with the d' values showing the same trend. The reason may be that 24mAs/120mAs had a dose ratio of five, while LD/HD-100% had a dose ratio of four and contained scatter and detector noises. A denoiser trained with a higher dose ratio or a less noisy target produced a smoother background, thus larger CNR values. For 24mAs/120mAs and LD/HD-100%, a few more percentages of MCs failed the Gaussian fitting than those in the LD images for the two smaller MC groups, indicating a greater loss of the relative subtle MCs, as also evident in Fig. 10. Fig. 11 shows that the CNRs of MCs in the DBCN images were comparable to those in the 24mAs/120mAs images but were sharper and had higher fit success rates. However, the backgrounds in the DBCN images in Fig. 10 appeared patchier and were noisier than the DNGAN images, which might have

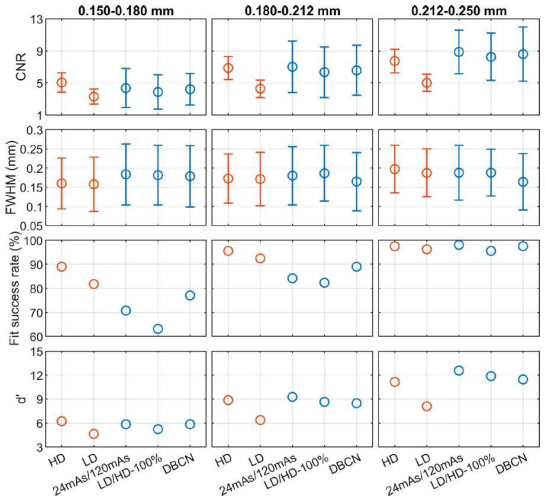


Fig. 11. The CNR, FWHM, fit success rates, and d' of the MCs in the validation physical phantom for comparing the denoising results.

contributed to the perceived noise and led to the lower d' of DBCN than those of 24mAs/120mAs and LD/HD-100% for the two larger MC groups. The CNRs of MCs in the DBCN images were significantly higher than those in the LD images ($p < 0.001$ for all three MC sizes).

F. Denoising Performance on Human Subject DBTs

We deployed the DNGAN denoisers (24mAs/120mAs and 24mAs/noiseless) to the human subject DBTs for independent testing. Fig. 12 shows that both denoisers and the DBCN were capable of reducing noise and maintaining the margins of the spiculated mass (invasive ductal carcinoma) and improving the conspicuity of the MC cluster (ductal carcinoma in situ). Although the background tissue of the 24mAs/noiseless denoised images was smooth as we discussed in Section III.B, the spiculations were still well preserved. The DBCN images had a patchy and noisier breast parenchyma than the DNGAN denoised images.

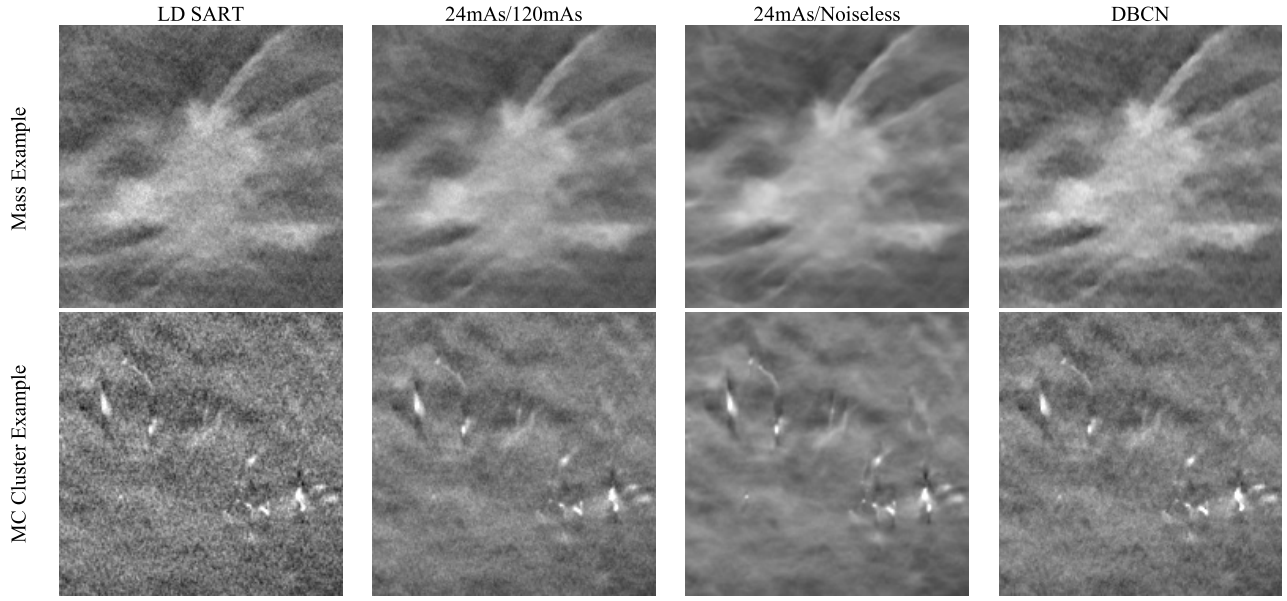


Fig. 12. Example images of human subject DBTs with a spiculated mass (invasive ductal carcinoma) and an MC cluster (ductal carcinoma in situ). All images show an 18 mm × 18 mm region. The images in the same row are displayed with the same window/level settings.

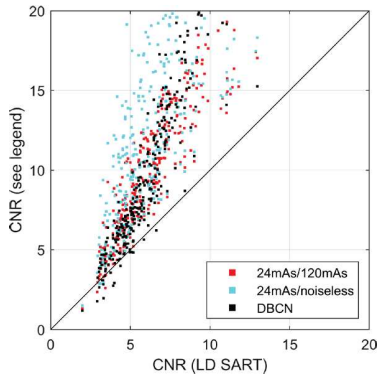


Fig. 13. The CNR scatter plot of MCs in the human subject DBTs for the DNGAN denoised images and the DBCN reconstructed images versus the LD SART images.

Because the MCs in human subjects did not have nominal sizes, instead of comparing the d' or average CNR values, we generated the CNR scatter plot of individual MCs, as shown in Fig. 13. The CNRs of most MCs were improved after DNGAN denoising. The CNRs of the 24mAs/120mAs denoised images were comparable to those of the DBCN images. The 24mAs/noiseless denoised images had the highest CNRs. However, whether the smooth appearance of the breast parenchyma is acceptable to radiologists and whether it has any effect on diagnosis will warrant future investigations.

G. Denoising Pristina-Reconstructed Images Using SART Denoiser

To evaluate the generalizability of our DNGAN denoisers in terms of the reconstruction algorithms, we deployed the denoisers trained with SART-reconstructed images to the LD validation physical phantom image that was reconstructed by the Pristina algorithm. Specifically, we selected the 24mAs/120mAs and LD/HD-100% denoisers that were used in Section III.E. We also trained a matched denoiser using the LD/HD-Pristina set.

Fig. 14 shows that the 24mAs/120mAs and LD/HD-100% denoisers worked to certain extent even though they were trained using SART-reconstructed images. The background texture of LD/HD-Pristina denoised images was visually more similar to that of the HD Pristina reference, whereas the 24mAs/120mAs and LD/HD-100% denoisers produced smoother appearance. Fig. 15 shows that all three denoisers reduced the noise and improved the CNRs significantly ($p < 0.001$ for all three MC sizes) compared to the LD Pristina-reconstructed images. LD/HD-Pristina had higher MC fit success rate and lower FWHM than the other two mismatched denoisers. Fig. 14 and Fig. 15 also included the results of DBCN from Fig. 10 and Fig. 11 to facilitate comparison.

IV. Discussion

The proposed DNGAN enjoys three aspects of robustness. First, the DNGAN trained with phantom data is applicable to human subject DBTs. This avoids the need to train using HD human DBTs, which may be impossible to collect. Second, the DNGAN can be trained with either digital phantom data or physical phantom data. This allows much flexibility in terms of the training data preparation. The digital phantom data have some advantages over the physical phantom data. For example, the software packages for producing the digital phantom data are open-source. It is inexpensive to generate a large set of data once the simulation model is formed, whereas making a large number of realistic physical phantoms is difficult. The high dose level of imaging a physical phantom is also limited by the tube loading of the DBT system. Third, the DNGAN trained with SART-reconstructed images is transferable to denoise other types of images such as Pristina-reconstructed images, although the denoising performance is not as good

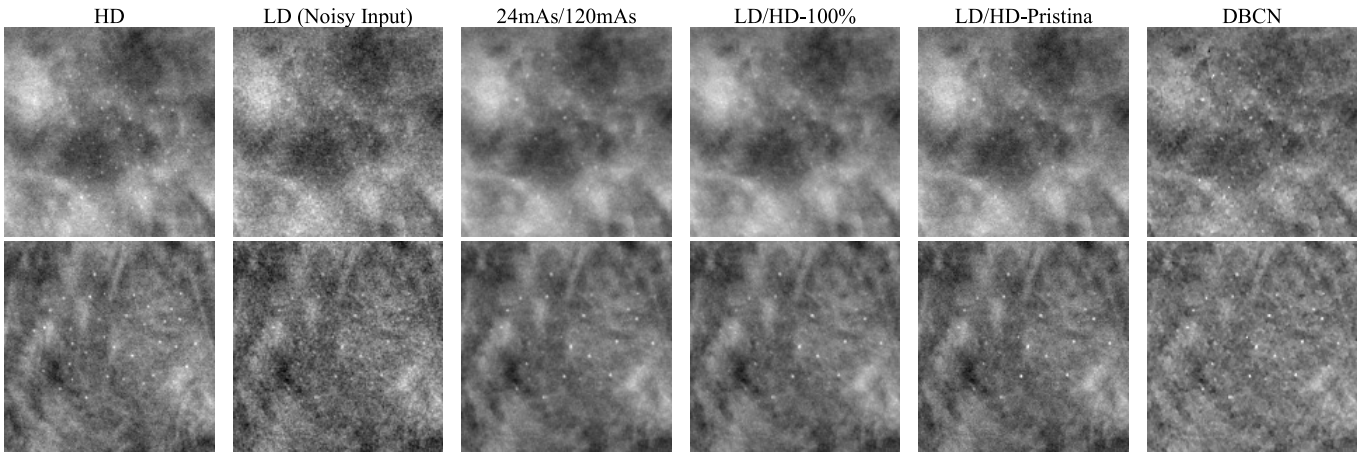


Fig. 14. Example MC clusters in the Pristina-reconstructed images of the validation physical phantom for comparing the mismatched and matched denoisers. Top row: 0.150-0.180 mm cluster. Bottom row: 0.180-0.212 mm cluster. All images show a 15 mm $\times$ 15 mm region. The images in the same row are displayed with the same window/level settings.

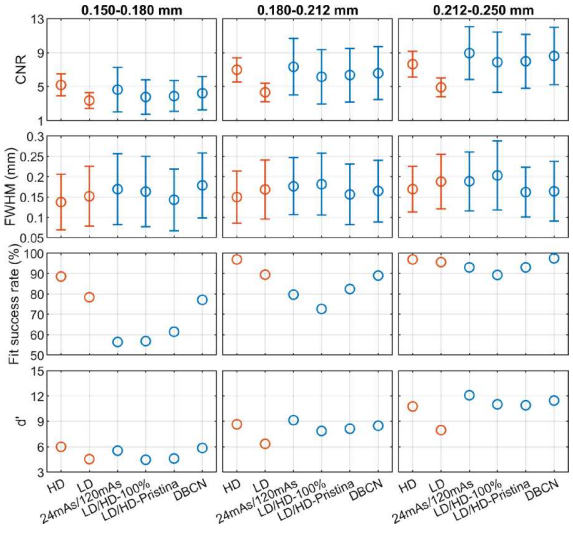


Fig. 15. The CNR, FWHM, fit success rates, and d' of the MCs in the Pristina-reconstructed images of the validation physical phantom for comparing the mismatched and matched denoisers.

as that obtained with a denoiser trained with data from matched reconstruction. This makes training DNGAN with *in silico* data applicable to clinical DBT images for which the reconstruction algorithm is proprietary such as those used in the commercial systems.

For DBT denoising, it seems to be less strict for the training data to be statistically representative of the patient population to achieve generalizability, as we demonstrated that the denoisers trained with digital phantoms were quite effective for physical phantoms and, most importantly, human subject DBTs. One explanation is that the mapping function for DBT denoising is simpler than those for predicting diseases or other clinical tasks. Nevertheless, we only tested the denoiser on a small set of human subject DBTs, so follow-up studies with DBTs of a wide range of properties are needed.

The DNGAN still smoothed out a substantial fraction of the very subtle MCs. We studied the feasibility of a second fine-tuning stage to improve the CNR and d' of subtle MCs, but the gain was offset by the increased spurious enhancement of noise and background structures (Section III.C). Using our current fine-tuning approach within the DNGAN framework, we have not found a good training condition that could balance between MC enhancement and spurious noise suppression. Further investigations of the training framework to enable the denoiser to distinguish MCs more effectively from noise and selectively enhance the true MCs are warranted.

We compared the image quality obtained from the proposed DNGAN denoising and our DBCN reconstruction. The two approaches represent two different directions to enhance the subtle signals in DBT. The DBCN models the detector blur and correlated noise of the imaging system, which was simplified to essentially a high-frequency-boosting filter on the PVs. To control the high-frequency noise, the DBCN was implemented with an edge-preserving regularizer. However, the reconstructed image quality was sensitive to the choice of the parameters of the regularizer and improper parameters may cause patchy soft tissue texture as discussed by Zheng *et al*. [10]. In contrast, the DNGAN smoothed the background around the signals to improve their conspicuity, similar to the role of a regularizer, but it also smoothed out some subtle MCs. It would be interesting to incorporate the state-of-the-art deep learning denoiser and the DBCN model into one reconstruction framework [57], [58]. Another noteworthy difference between the two methods is that DNGAN is a post-processing approach, whereas DBCN is a reconstruction algorithm. The DNGAN training is relatively flexible and, once fully trained, the DNGAN denoiser is readily deployable to the reconstructed DBT images and potentially applicable to DBT from different reconstruction techniques, as demonstrated in our study. In contrast, DBCN models a given DBT imaging system and requires the raw PVs that may not be stored or accessible in clinical practice. Both approaches have their advantages and disadvantages, and the choice will require future studies to compare the overall cancer detection accuracy and assess the preference of the image appearance by radiologists.

We observed a good correlation between CNR and d' that we calculated to assess the conspicuity of MCs. Fig. 16 shows

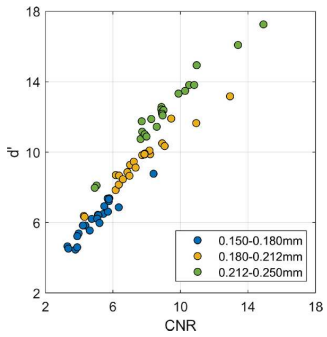


Fig. 16. Scatter plot of d' versus CNR, including all data points from Fig. 3, 5, 8, 11, and 15 for three nominal MC sizes and different conditions.

a scatter plot of the mean CNR and the corresponding d' from the results in Section III. The Spearman rank correlation coefficient between CNR and d' was $\rho = 0.96$ and the correlation was statistically significant ($p < 0.001$). This observation suggests that the simple CNR might be a good surrogate for the more sophisticated d' as an image quality metric of MCs for the task in this study.

The figures of merit we used in this study have their limitations. First, CNR was used as an indicator for the conspicuity of individual MCs, but we only calculated CNR at known locations. The clinical usefulness of an image enhancement method has to consider both the true signals and the falsely enhanced noise or structures in an image. Alternative methods, such as computerized detection [59], can be used to study the tradeoff between the increase in detectability of MCs and false positive detection in the future. Second, NPS and d' are Fourier-based, but the DCNN denoiser is nonlinear and DBT is shift-variant. Third, NPS provides a relative ranking of the noise level of the images, but it does not reflect the visual quality of the soft tissues or masses. To our knowledge, there is no figure of merit available to describe the fine textural appearance of an image or a soft-tissue lesion and correlate it with human visual preference. This makes it difficult to objectively optimize the balance between image smoothness and MC enhancement (Section III.B and Section III.F). The acceptability of the image quality or image appearance for clinical reading will have to be judged by radiologists in human subject DBTs. Reader studies with radiologists can provide more clinically relevant assessments about the pros and cons of each condition but it is impractical to conduct reader studies for many conditions because of the limited availability of radiologists' time. Reader studies are beyond the scope of this paper.

V. Conclusion

We developed a DNGAN framework based on adversarial training for denoising reconstructed DBT images. A properly weighted combination of an MSE training loss and an adversarial loss was found to be effective for noise reduction and texture preservation. We demonstrated the impacts of the dose level of the training targets and the training sample size on the performance of DBT denoising. We evaluated a fine-tuning stage to further enhance subtle MCs but found that it also enhanced false positives and was unsuitable for practical use. The DNGAN could be trained using *in silico* data and applied to physical phantom images even from a different reconstruction algorithm. Promising preliminary results were observed in deploying the trained denoiser to a small set of human subject DBTs. Further studies will be conducted to evaluate the effects of the denoiser on the detection of MCs and subtle lesions in DBT by computer vision or human readers.

Acknowledgment

The authors would like to thank Ravi Samala, Ph.D., for the helpful discussions about DCNN training, Jiabei Zheng, Ph.D., for providing the MTF, NPS, CNR and DBCN programs, GE Global Research for providing the CatSim simulation tool, and Christian Graff, Ph.D., for the open-source digital breast phantom generation program.

References

[1] J. T. Dobbins, "Tomosynthesis imaging: At a translational crossroads," *Med. Phys.*, vol. 36, no. 6Part1, pp. 1956–1967, May 2009.
[2] I. Sechopoulos, "A review of breast tomosynthesis. Part I. The image acquisition process," *Med. Phys.*, vol. 40, no. 1, Jan. 2013, Art. no. 014301.
[3] H.-P. Chan *et al.*, "Digital breast tomosynthesis: Observer performance of clustered microcalcification detection on breast phantom images acquired with an experimental system using variable scan angles, angular increments, and number of projection views," *Radiology*, vol. 273, no. 3, pp. 675–685, Dec. 2014.
[4] M. M. Goodsitt *et al.*, "Digital breast tomosynthesis: Studies of the effects of acquisition geometry on contrast-to-noise ratio and observer preference of low-contrast objects in breast phantom images," *Phys. Med. Biol.*, vol. 59, no. 19, pp. 5883–5902, Oct. 2014.
[5] J. Ludwig, T. Mertelmeier, H. Kunze, and W. Härer, "A novel approach for filtered backprojection in tomosynthesis based on filter kernels determined by iterative reconstruction techniques," in *Proc. Digit. Mammography (IWDM)*, vol. 5116, 2008, pp. 612–620.
[6] J. Liu *et al.*, "Radiation dose reduction in digital breast tomosynthesis (DBT) by means of deep-learning-based supervised image processing," *Proc. SPIE*, vol. 10574, Mar. 2018, Art. no. 105740F.
[7] Y. Zhang *et al.*, "A comparative study of limited-angle cone-beam reconstruction methods for breast tomosynthesis," *Med. Phys.*, vol. 33, no. 10, pp. 3781–3795, Oct. 2006.
[8] M. Das, H. C. Gifford, J. M. O'Connor, and S. J. Glick, "Penalized maximum likelihood reconstruction for improved microcalcification detection in breast tomosynthesis," *IEEE Trans. Med. Imag.*, vol. 30, no. 4, pp. 904–914, Apr. 2011.
[9] S. Xu, J. Lu, O. Zhou, and Y. Chen, "Statistical iterative reconstruction to improve image quality for digital breast tomosynthesis," *Med. Phys.*, vol. 42, no. 9, pp. 5377–5390, Sep. 2015.
[10] J. Zheng, J. A. Fessler, and H.-P. Chan, "Detector blur and correlated noise modeling for digital breast tomosynthesis reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 116–127, Jan. 2018.
[11] K. Kim *et al.*, "Fully iterative scatter corrected digital breast tomosynthesis using GPU-based fast Monte Carlo simulation and composition ratio update," *Med. Phys.*, vol. 42, no. 9, pp. 5342–5355, Aug. 2015.
[12] Y. Lu, H.-P. Chan, J. Wei, and L. M. Hadjiiski, "Selective-diffusion regularization for enhancement of microcalcifications in digital breast tomosynthesis reconstruction," *Med. Phys.*, vol. 37, no. 11, pp. 6003–6014, Nov. 2010.
[13] E. Y. Sidky, I. S. Reiser, R. Nishikawa, and X. Pan, "Image reconstruction in digital breast tomosynthesis by total variation minimization," *Proc. SPIE*, vol. 6510, Mar. 2007, Art. no. 651027.
[14] I. Kastanis *et al.*, "3D digital breast tomosynthesis using total variation regularization," in *Proc. Digit. Mammography (IWDM)*, vol. 5116, 2008, pp. 621–627.
[15] E. Y. Sidky, X. Pan, I. S. Reiser, R. M. Nishikawa, R. H. Moore, and D. B. Kopans, "Enhanced imaging of microcalcifications in digital breast tomosynthesis through improved image-reconstruction algorithms," *Med. Phys.*, vol. 36, no. 11, pp. 4920–4932, Nov. 2009.
[16] J. Zheng, J. A. Fessler, and H.-P. Chan, "Digital breast tomosynthesis reconstruction using spatially weighted non-convex regularization," *Proc. SPIE*, vol. 9783, Mar. 2016, Art. no. 978369.

[17] M. Sghaier, E. Chouzenoux, J. Pesquet, and S. Muller, "A new spatially adaptive TV regularization for digital breast tomosynthesis," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2020, pp. 629–633.

[18] F. E. Boas and D. Fleischmann, "CT artifacts: Causes and reduction techniques," *Imag. Med.*, vol. 4, no. 2, pp. 229–240, Apr. 2012.

[19] S. Kligerman *et al.*, "Detection of pulmonary embolism on computed tomography: Improvement using a model-based iterative reconstruction algorithm compared with filtered back projection and iterative reconstruction algorithms," *J. Thoracic Imag.*, vol. 30, no. 1, pp. 60–68, Jan. 2015.

[20] M. Das, C. Connolly, S. J. Glick, and H. C. Gifford, "Effect of postreconstruction filter strength on microcalcification detection at different imaging doses in digital breast tomosynthesis: Human and model observer studies," *Proc. SPIE*, vol. 8313, Mar. 2012, Art. no. 831321.

[21] S. Abdurahman, F. Dennerlein, A. Jerebko, A. Fieselmann, and T. Mertelmeier, "Optimizing high resolution reconstruction in digital breast tomosynthesis using filtered back projection," in *Digit. Mammography (IWDM)*, vol. 8539, 2014, pp. 520–527.

[22] Y. Lu, H.-P. Chan, J. Wei, L. M. Hadjiiski, and R. K. Samala, "Multiscale bilateral filtering for improving image quality in digital breast tomosynthesis," *Med. Phys.*, vol. 42, no. 1, pp. 182–195, Jan. 2015.

[23] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[26] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[28] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*. [Online]. Available: http://arxiv.org/abs/1406.2661

[29] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," 2017, *arXiv:1704.00028*. [Online]. Available: http://arxiv.org/abs/1704.00028

[32] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017.

[33] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[34] Z. Hu *et al.*, "Artifact correction in low-dose dental CT imaging using Wasserstein generative adversarial networks," *Med. Phys.*, vol. 46, no. 4, pp. 1686–1696, Apr. 2019.

[35] H. Shan *et al.*, "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1522–1534, Jun. 2018.

[36] H. Shan *et al.*, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," *Nature Mach. Intell.*, vol. 1, no. 6, pp. 269–276, Jun. 2019.

[37] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[38] M. Mardani *et al.*, "Deep generative adversarial neural networks for compressive sensing MRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 167–179, Jan. 2019.

[39] M. Gao, R. K. Samala, J. A. Fessler, and H.-P. Chan, "Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction," *Proc. SPIE*, vol. 11312, Mar. 2020, Art. no. 113120Q.

[40] A. Badano *et al.*, "Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial," *JAMA Netw. Open*, vol. 1, no. 7, Nov. 2018, Art. no. e185474.

[41] C. G. Graff, "A new, open-source, multi-modality digital breast phantom," *Proc. SPIE*, vol. 9783, Mar. 2016, Art. no. 978309.

[42] B. De Man *et al.*, "CatSim: A new computer assisted tomography simulation environment," *Proc. SPIE*, vol. 6510, Mar. 2007, Art. no. 65102G.

[43] B. De Man, J. Pack, P. FitzGerald, and M. Wu, "CatSim manual version 6.0," GE Global Res., Niskayuna, NY, USA, Tech. Rep., 2015.

[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5967–5976.

[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" 2014, *arXiv:1411.1792*. [Online]. Available: http://arxiv.org/abs/1411.1792

[46] Public Health England. (2019). *Technical Evaluation of GE Healthcare Senographe Pristina Digital Breast Tomosynthesis System*. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/771830/GE_Pristina_Tomo.pdf

[47] J. Zheng, J. A. Fessler, and H.-P. Chan, "Segmented separable footprint projector for digital breast tomosynthesis and its application for subpixel reconstruction," *Med. Phys.*, vol. 44, no. 3, pp. 986–1001, Mar. 2017.

[48] R. K. Samala *et al.*, "Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 686–696, Mar. 2019.

[49] A. E. Burgess, "Mammographic structure: Data preparation and spatial statistics analysis," *Proc. SPIE*, vol. 3661, pp. 642–653, Mar. 1999.

[50] A. E. Burgess, "Statistically defined backgrounds:Performance of a modified nonprewhitening observer model," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 11, no. 4, p. 1237, Apr. 1994.

[51] L.-N. D. Loo, K. Doi, and C. E. Metz, "A comparison of physical image quality indices and observer performance in the radiographic detection of nylon beads," *Phys. Med. Biol.*, vol. 29, no. 7, pp. 837–856, Jul. 1984.

[52] A. E. Burgess, X. Li, and C. K. Abbey, "Visual signal detectability with two noise components: Anomalous masking effects," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 14, no. 9, p. 2420, Sep. 1997.

[53] S. Richard and J. H. Siewerdsen, "Comparison of model and human observer performance for detection and discrimination tasks using dual-energy X-ray images," *Med. Phys.*, vol. 35, no. 11, pp. 5043–5053, Oct. 2008.

[54] G. J. Gang *et al.*, "Analysis of Fourier-domain task-based detectability index in tomosynthesis and cone-beam CT in relation to human observer performance," *Med. Phys.*, vol. 38, no. 4, pp. 1754–1768, Mar. 2011.

[55] O. Christianson *et al.*, "An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations," *Radiology*, vol. 275, no. 3, pp. 725–734, Jun. 2015.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[57] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM J. Imag. Sci.*, vol. 10, no. 4, pp. 1804–1844, Jan. 2017.

[58] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, Mar. 2017.

[59] R. K. Samala, H.-P. Chan, Y. Lu, L. M. Hadjiiski, J. Wei, and M. A. Helvie, "Computer-aided detection system for clustered microcalcifications in digital breast tomosynthesis using joint information from volumetric and planar projection images," *Phys. Med. Biol.*, vol. 60, no. 21, pp. 8457–8479, Nov. 2015.

Deep Convolutional Neural Network with Adversarial Training for Denoising Digital Breast Tomosynthesis Images

Supplementary Materials

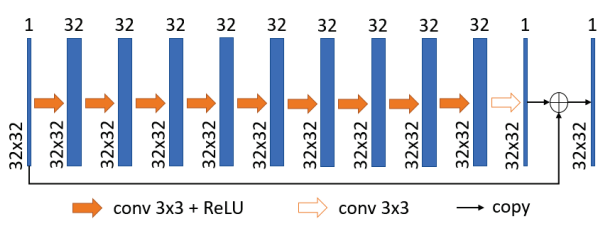
Mingjie Gao, *Graduate Student Member, IEEE*, Jeffrey A. Fessler, *Fellow, IEEE*, and Heang-Ping Chan

3/17/2021

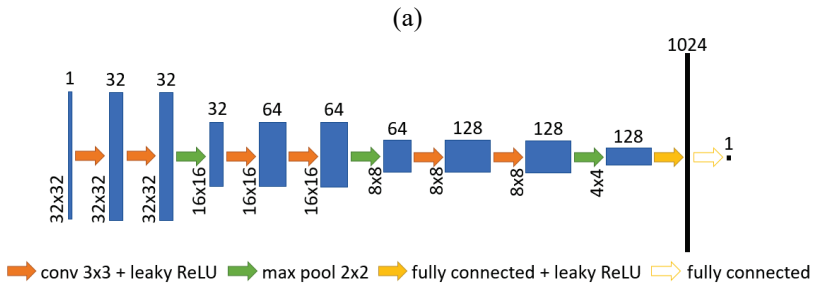
S-I. Network structures of denoiser and discriminator

We designed the network structure of the denoiser based on the DnCNN [1]. The original DnCNN structure consisted of 17 to 20 convolutional layers, each included 64 filters of 3×3 kernels, and used rectified linear units (ReLU) between the layers. Each convolution layer was followed by batch normalization [2] before the ReLU except for the first convolution layer. For our DBT denoising task, we chose to use LD/HD image regions, or patches, 32×32 pixels in size as inputs to allow the DCNN to focus on the local image structures in the adversarial training [3]. Our pilot studies found that the structure could be reduced to 10 convolutional layers, each with 32 filters, without substantial difference in performance. We also removed the batch normalization layers without experiencing training instability [4]. We therefore used a much smaller structure, as shown in Figure S1(a), to improve computational efficiency. This structure had a total of 74,593 trainable weights. To ensure that the output of the convolution layer had the same size as the input, the input was padded with values that were mirrored from the inner region of the input. Because the denoiser was fully convolutional, we could directly apply it to the full DBT slices during deployment.

Figure S1(b) shows the network structure of the discriminator. We used the VGG-Net [5], with a reduced number of downsampling blocks due to the small input patch size, as the backbone of our discriminator. The discriminator had a total of 2,385,633 trainable weights.



(a)



(b)

Figure S1. The network structure of (a) the denoiser and (b) the discriminator. The numbers on the left of each layer (rectangle) represent the width and height of the tensors. The numbers on the top of the rectangles represent the number convolution filters for the tensors (blue) or the length of the vectors (black).

S-II. Derivation of the adversarial loss

We implemented the adversarial loss as the WGAN with gradient penalty [6]. The key idea behind it is summarized as follows. Assume $\Omega = \mathbb{R}^{N_{\text{pixel}}}$ is the sample space of DBT images, $\text{Prob}(\Omega)$ is the space of probability measures defined on Ω. We can view the denoiser $G: \Omega \to \Omega$ as a function parameterized by its trainable weights. It generates denoised images following a distribution $\mathbb{P}_G \in \text{Prob}(\Omega)$ from the input samples that follow a noisy image distribution $\mathbb{P}_{\text{noisy}} \in \text{Prob}(\Omega)$. The Wasserstein distance (WD) between the distribution of the denoised images $\mathbb{P}_G$ and the distribution of the HD target images $\mathbb{P}_{\text{target}} \in \text{Prob}(\Omega)$ is defined as [4]

$$d(\mathbb{P}_G, \mathbb{P}_{\text{target}}) = \inf_{\gamma \in \Pi(\mathbb{P}_G, \mathbb{P}_{\text{target}})} \mathbb{E}_{(p,q)\sim\gamma}[\|p - q\|] \tag{S1}$$

where $\Pi(\mathbb{P}_G, \mathbb{P}_{\text{target}})$ denotes the set of all joint distributions $\gamma(p,q)$ whose marginals are $\mathbb{P}_G$ and $\mathbb{P}_{\text{target}}$, respectively. Arjovsky *et al*. [4] showed that, instead of directly solving (S1), which is intractable, one could solve

$$d(\mathbb{P}_G, \mathbb{P}_{\text{target}}) = \max_{\|D\|_L \le 1} \mathbb{E}_{x_{\text{target}}}[D(x_{\text{target}})] - \mathbb{E}_{x_{\text{noisy}}}\left[D\left(G(x_{\text{noisy}})\right)\right]. \tag{S2}$$

In other words, to calculate the WD, we need to find a 1-Lipschitz function $D: \Omega \to \mathbb{R}$ that maximizes the objective function. D is also called a discriminator, or a critic, to output a similarity score that assesses whether the input image patch comes from the target distribution. The discriminator is approximated by a DCNN in our implementation. Its training loss function is [4]

$$\underset{\|D\|_L \le 1}{\text{argmax}}\, \hat{d}_G(D) \coloneqq \mathbb{E}_{x_{\text{target}}}[D(x_{\text{target}})] - \mathbb{E}_{x_{\text{noisy}}}\left[D\left(G(x_{\text{noisy}})\right)\right]. \tag{S3}$$

Arjovsky *et al*. showed that $\hat{d}_G(D)$ can be interpreted as an estimation of the WD. Note that $\hat{d}_G(D)$ also depends on G. Gulrajani *et al*. proposed a gradient penalty to constrain the 1-Lipschitz condition [6], so the overall training loss function for the discriminator becomes

$$\underset{D}{\text{argmin}} -\hat{d}_G(D) + \lambda_D \cdot \mathbb{E}_{\bar{x}}[(\|\nabla_{\bar{x}} D(\bar{x})\| - 1)^2] \tag{S4}$$

where $\bar{x} = t \cdot x_{\text{target}} + (1-t) \cdot G(x_{\text{noisy}})$ is an interpolated image, $t \sim \text{Unif}([0,1])$, λ_D is the penalty weight. After estimating WD, we expect the term $\mathbb{E}_{x_{\text{noisy}}}\left[D\left(G(x_{\text{noisy}})\right)\right]$ to be large when optimizing G to promote the denoised images to be perceptually similar to the target images. This gives the corresponding adversarial loss of the denoiser training

$$L_{\text{adv}}(G) = -\mathbb{E}_{x_{\text{noisy}}}\left[D\left(G(x_{\text{noisy}})\right)\right]. \tag{S5}$$

In practice, the denoiser and the discriminator are trained alternately in DNGAN so that the discriminator is always up to date for estimating $d(\mathbb{P}_G, \mathbb{P}_{\text{target}})$ and the denoiser improves through iterations.

S-III. CatSim configuration

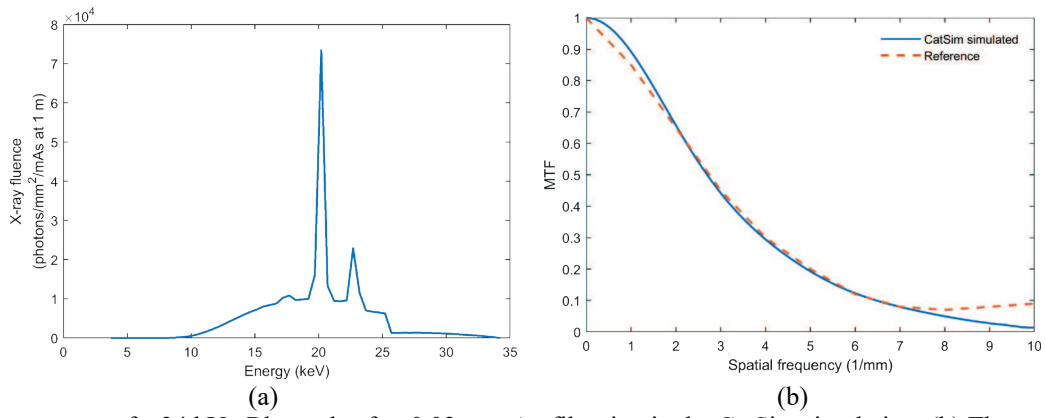
We configured CatSim [7][8] to model the GE Pristina DBT system (GE Healthcare) [9] as follows: set the acquisition geometry as 9 PVs within $\pm 12.5°$ at a detector pixel size of 0.1 mm × 0.1 mm; set the x-ray fluence spectrum at 34 kVp from a Rh anode [10] with a 0.03 mm Ag filter (Figure S2(a)); used the x-ray detection model developed by Carvalho [11] for the CsI/Si flat panel indirect detector. The simulated signal at the ith detector pixel $Y(i)$ is

$$Y(i) = c \cdot h_{\text{scint}} * \left(\sum_{E \in \mathcal{E}} E \cdot \text{Poisson}\{\eta(E,i) \cdot I_{\text{inc}}(E,i)\} \right) \tag{S6}$$

This work was supported by the National Institutes of Health Award Number R01 CA214981. M. Gao and J. A. Fessler are with the Department of Electrical Engineering and Computer Science and the Department of Radiology, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: gmingjie@umich.edu, fessler@umich.edu). H.-P. Chan is with the Department of Radiology, University of Michigan (e-mail: chanhp@umich.edu).

where c is a conversion factor from photons to electrons, h_{scint} is the scintillator blur kernel, $*$ represents spatial convolution, $\mathcal{E}$ is the set of energy bins in the input x-ray spectrum, Poisson$\{\cdot\}$ denotes Poisson distribution, $\eta(E,i)$ is the energy-dependent detection efficiency at detector element i, $I_{\text{inc}}(E,i)$ is the x-ray intensity spectrum incident on the detector element i. Focal-spot blur, scattered radiation and electronic noise were not considered in the simulation for this study. To validate the system response of CatSim simulation, we calculated its presampled modulation transfer function (MTF) for the central PV using the edge method [12]. It agreed well with the measured Pristina MTF in the literature [13], as shown in Figure S2(b).

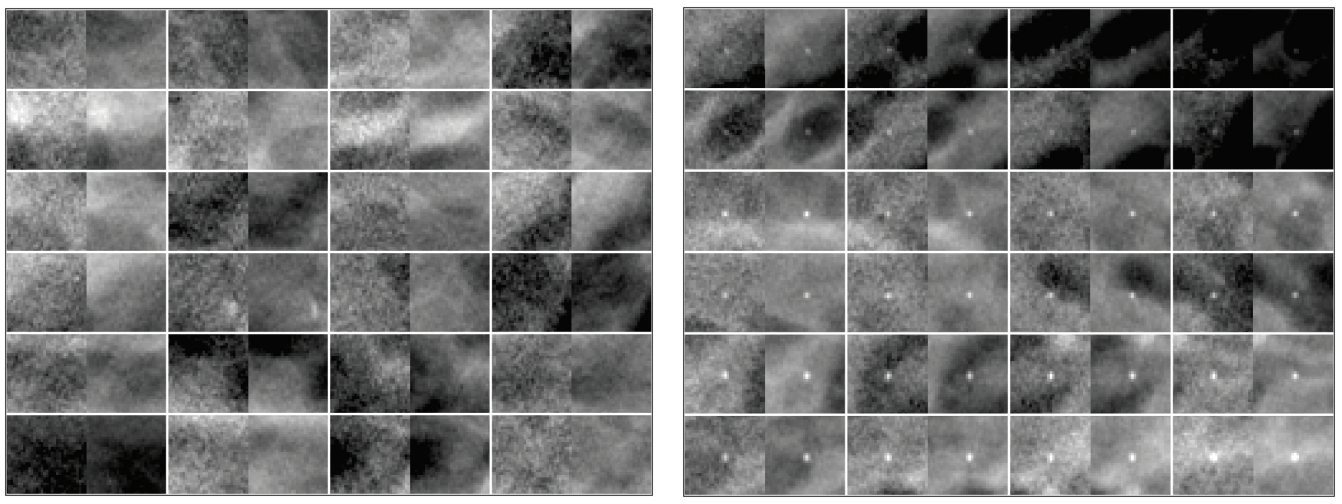
We set the total x-ray exposure of 9 PVs to 24 mAs for the 4.5-cm-thick VICTRE phantoms in CatSim. The estimated mean glandular dose (MGD) was 1.42 mGy under this exposure, calculated by a Monte Carlo simulation tool called CatDose in the CatSim package.



(a) (b)

Figure S2. (a) The x-ray fluence spectrum of a 34 kVp Rh anode after 0.03 mm Ag filtration in the CatSim simulation. (b) The system MTF for the central PV. The edge for MTF calculation was placed on the breast support plate and was parallel to the chest wall.

S-IV. Example training patches



(a) (b)

Figure S3. Examples of 24 pairs of (a) DBT patches in the 24mAs/120mAs training set and (b) MC patches in the accompanying MC fine-tuning set. In each pair, the LD patch is shown on the left and the HD patch on the right.

S-V. Figures of merit

Noise power spectrum (NPS)

For the noise power spectrum (NPS) calculation, we first extracted 40 background slice patches, 200×200 pixels each, from the DBT slices parallel to the detector plane, and then calculated the 2D NPS defined as [14]

$$\text{NPS}_{\text{2D}} = \frac{p_x p_y}{N_x N_y} \langle |\text{DFT}_{\text{2D}}\{x_i - \bar{x}_i\}|^2 \rangle_i \tag{S7}$$

where $p_x = p_y = 0.1$ mm is the image pixel size, $N_x = N_y = 200$ is the patch size, $\langle \cdot \rangle_i$ means averaging over all patches, $\text{DFT}_{\text{2D}}\{\cdot\}$ denotes 2D discrete Fourier transform, x_i is the image patch, $\bar{x}_i$ is the mean pixel value of the patch, $i = 1, \dots, 40$. Finally, the 1D NPS was calculated by taking the rotational average of the 2D NPS.

Contrast-to-noise ratio (CNR) and full width at half maximum (FWHM)

To quantitatively evaluate the MCs in the images, we calculated the contrast-to-noise ratio (CNR) and full width at half maximum (FWHM) as figures of merit for each MC. CNR indicates the conspicuity of MCs within the local surroundings, while FWHM measures their sharpness. Given a 32×32 patch with an MC at the center, we used a 2D Gaussian plus a 2D first order plane as the fitting function to fit the signal and the background in the central 13×13-pixel region. We define

$$\text{CNR} = \frac{I_{\text{MC}}}{\sigma_{\text{bg}}}, \ \text{FWHM} = 2\sqrt{2\ln 2} \cdot \sigma_{\text{MC}} \tag{S8}$$

where I_{MC} is the maximum value of the 2D fitted Gaussian on the pixel grids, σ_{bg} is the root-mean-square noise of the surrounding area after removing the local background mean gray level using a box-rim filter [15] and excluding the MC pixels, σ_{MC} is the standard deviation of the fitted Gaussian. Figure S4 illustrates the workflow of the CNR and FWHM calculation.

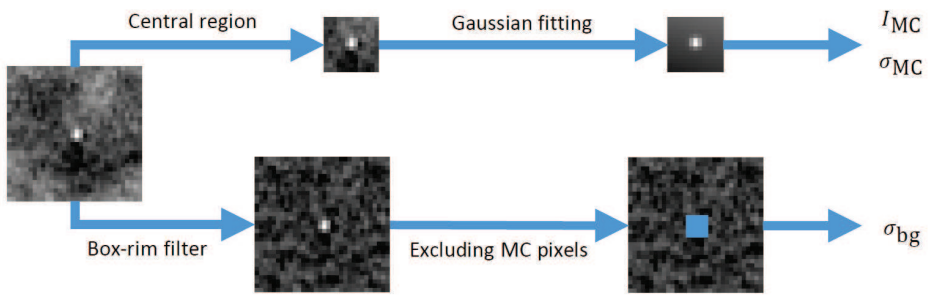


Figure S4. An illustration of CNR and FWHM calculation for an MC patch.

Detectability index (d')

We calculated the detectability index (d') using the nonprewhitening matched filter model observer with eye filter (NPWE) [16][17] for the MC detection task. We considered the 2D in-plane d' of the DBT slices in the validation physical phantom for each of the study conditions

$$d' = \frac{\iint S^2(u,v)\, E^2(u,v)\, du\, dv}{\left(\iint W(u,v)\, S^2(u,v)\, E^4(u,v)\, du\, dv\right)^{1/2}} \tag{S9}$$

where u, v are spatial frequencies in 1/mm, S is the (blurred) signal spectrum (the product of the task function and the task transfer function), W is the 2D NPS, E is the eye filter or the visual response function of a human observer.

In our study, the imaging task was to detect MCs of different nominal sizes in the heterogeneous background of breast phantom DBT. Similar to the CNR calculation, we assumed the MC shape to be Gaussian. We calculated the average d' for each MC speck size group by using the averages of the fitted parameters, contrast $\bar{I}_{\mathrm{MC}}$ and standard deviation $\bar{\sigma}_{\mathrm{MC}}$, obtained from the Gaussian fitting to the individual MCs. The signal spectrum was thus given by

$$s(x,y) = \bar{I}_{\mathrm{MC}} \cdot \exp\left(-\frac{x^2+y^2}{2\bar{\sigma}_{\mathrm{MC}}^2}\right) \xleftrightarrow{\text{Fourier transform}} S(u,v) = \bar{I}_{\mathrm{MC}} \cdot 2\pi\bar{\sigma}_{\mathrm{MC}}^2 \cdot \exp\left(-2\pi^2\bar{\sigma}_{\mathrm{MC}}^2(u^2+v^2)\right). \tag{S10}$$

The NPS was calculated by (S7). It characterized the structured noise of the image background as well as other noise from the imaging chain. We used the theoretical model of the eye filter that was proposed by Kelly [18]

$$E(u,v) = (u^2+v^2) \cdot \exp\left(-c\sqrt{u^2+v^2}\right). \tag{S11}$$

Considering that radiologists usually search for MCs in magnification mode because of their small size, we set the viewing distance to 12.5 cm, which corresponds to 4 times higher magnification than the usual 50 cm viewing distance. Under this condition, the value of c was set to 1 so that the eye filter had its peak at 4 cycles/deg [16].

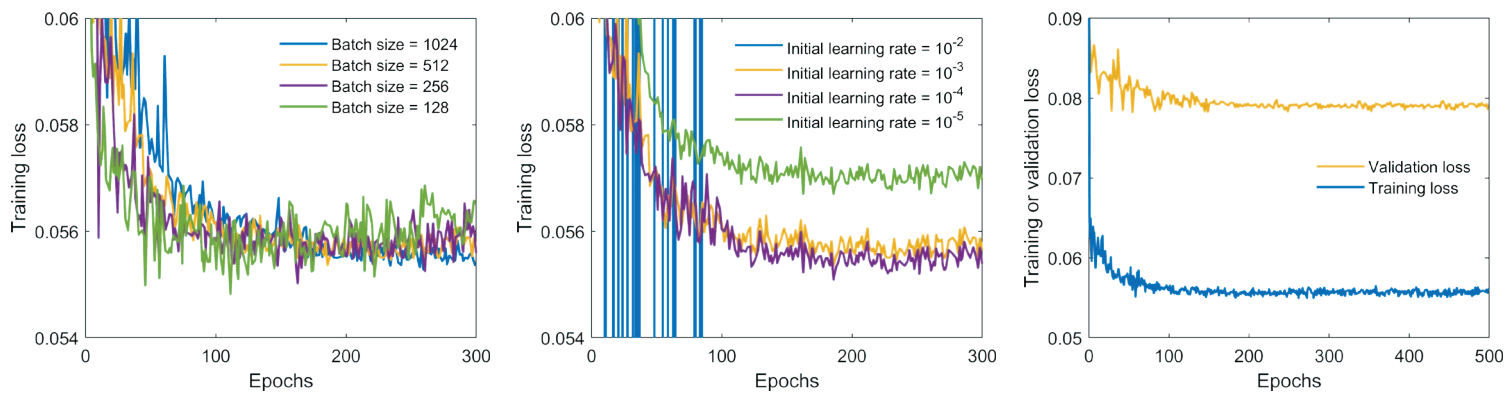
S-VI. Effects of training parameters

We studied the effects of several training parameters including the batch size, the learning rate, and the number of epochs. The training set was the 24mAs/120mAs digital phantom set. If not specified, other training parameters were the same as those described in the paper. We used the same random seeds for weight initialization and data batching for all conditions.

To study the effect of the batch size, we trained four denoisers using the batch sizes of 1024, 512, 256, and 128 while keeping all other training parameters the same. Figure S5(a) shows the training losses (Eq. (1) in paper) of the four conditions. For a small batch size such as 128, the training loss decreased more quickly in early epochs than others because the weights were updated more frequently. However, in later epochs, the gradient estimated from a small batch had larger variations than that estimated from a larger batch, so the training loss had a large oscillation. For a large batch size such as 1024, the training occupied more GPU memory and also converged more slowly than the others. In our study, we used the batch size of 512.

To study the effect of the learning rate, we trained four denoisers using the initial learning rates of 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5} while keeping all other training parameters the same. Figure S5(b) shows the training losses of the four conditions. For a large initial learning rate such as 10^{-2}, the training oscillated and did not converge. For a small initial learning rate such as 10^{-5}, the training step size was small, so the training might be trapped by local minima. In our study, we used the initial learning rate of 10^{-3}.

To study the effect of the number of epochs, we ran the denoiser training up to 500 epochs. We also prepared a validation set to monitor if overfitting occurred in our training. The validation set had 53,141 paired patches that were extracted from the LD/HD validation physical phantom pair in the same way as for the training set. Figure S5(c) shows the training and validation losses. We did not observe overfitting because both the training and validation losses were stable and the gap between them remained approximately constant after about 200 epochs. We determined that 300 epochs were sufficient to achieve training convergence. The stability and convergence of the training loss achieved within 300 epochs were also observed for other conditions. More importantly, the robustness of many of the trained denoisers was validated by their across-phantom performance in the independent validation set (i.e., digital-phantom-data-trained denoiser applied to physical phantom images) and further on the independent unseen test set of human subject DBT images.



(a) (b) (c)

Figure S5. The training losses versus epochs for illustrating the effects of (a) the batch size, (b) the initial learning rate, and (c) the number of epochs.

References

[1] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, DOI: 10.1109/TIP.2017.2662206.

[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: http://arxiv.org/abs/1502.03167.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976, DOI: 10.1109/CVPR.2017.632.

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 26-Jan-2017. [Online]. Available: http://arxiv.org/abs/1701.07875.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 31-Mar-2017. [Online]. Available: http://arxiv.org/abs/1704.00028.

[7] B. De Man *et al.*, "CatSim: a new Computer Assisted Tomography SIMulation environment," in *Proceedings of SPIE*, 2007, p. 65102G, DOI: 10.1117/12.710713.

[8] B. De Man, J. Pack, P. FitzGerald, and M. Wu, "CatSim manual version 6.0," *GE Global Research*. 2015.

[9] "GE Senographe Pristina operator manual," *General Electric Company*, 2019. [Online]. Available: https://customer-doc.cloud.gehealthcare.com/.

[10] J. M. Boone, T. R. Fewell, and R. J. Jennings, "Molybdenum, rhodium, and tungsten anode spectral models using interpolating polynomials with application to mammography," *Med. Phys.*, vol. 24, no. 12, pp. 1863–1874, Dec. 1997, DOI: 10.1118/1.598100.

[11] P. Milioni de Carvalho, "Low-dose 3D quantitative vascular x-ray imaging of the breast," Université Paris Sud, 2014.

[12] E. Samei, M. J. Flynn, and D. A. Reimann, "A method for measuring the presampled MTF of digital radiographic systems using an edge test device," *Med. Phys.*, vol. 25, no. 1, pp. 102–113, Jan. 1998, DOI: 10.1118/1.598165.

[13] "Technical evaluation of GE Healthcare Senographe Pristina digital breast tomosynthesis system," *Public Health England*, 2019. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/771830/GE_Pristina_Tomo.pdf.

[14] J. H. Siewerdsen, I. A. Cunningham, and D. A. Jaffray, "A framework for noise-power spectrum analysis of multidimensional images," *Med. Phys.*, vol. 29, no. 11, pp. 2655–2671, Nov. 2002, DOI: 10.1118/1.1513158.

[15] B. Sahiner *et al.*, "Computer-aided detection of clustered microcalcifications in digital breast tomosynthesis: a 3D approach," *Med. Phys.*, vol. 39, no. 1, pp. 28–39, Jan. 2012, DOI: 10.1118/1.3662072.

[16] A. E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening observer model," *J. Opt. Soc. Am. A*, vol. 11, no. 4, p. 1237, Apr. 1994, DOI: 10.1364/JOSAA.11.001237.

[17] L.-N. D. Loo, K. Doi, and C. E. Metz, "A comparison of physical image quality indices and observer performance in the radiographic detection of nylon beads," *Phys. Med. Biol.*, vol. 29, no. 7, pp. 837–856, Jul. 1984, DOI: 10.1088/0031-9155/29/7/007.

[18] D. H. Kelly, "Spatial frequency selectivity in the retina," *Vision Res.*, vol. 15, no. 6, pp. 665–672, Jun. 1975, DOI: 10.1016/0042-6989(75)90282-5.