Generalizability performance of a deep learning-based CT image denoising method

Rongping Zeng, Claire Yilin Lin, Qin Li, Jiang Lu, Marlene Skopec, Jeffrey A Fessler and Kyle J Myers

ABSTRACT

Purpose: Deep learning (DL) is rapidly finding applications in low-dose CT image denoising. While having the potential to improve image quality over the filtered back projection method (FBP) and produce images quickly, performance generalizability of the data-driven DL methods is not fully understood yet. The main purpose of this work is to investigate the performance generalizability of a low-dose CT image denoising neural network in data acquired under different scan conditions, particularly relating to these three parameters: reconstruction kernel, slice thickness and dose (noise) level. A secondary goal is to identify any underlying data property associated with the CT scan settings that might help predict the generalizability of the denoising network.

Methods: We select the residual encoder-decoder convolutional neural network (REDCNN) as an example of a low-dose CT image denoising technique in this work. We use the patient scans in the Low-Dose Grand Challenge (LDGC) dataset to train the network. To study how the network generalizes on the three acquisition parameters, we analyze the denoising performance changes under three scenarios: smooth vs sharp reconstruction kernels, 1 mm vs 3 mm slice thicknesses, fixed (25%) vs mixed dose levels. In each scenario, we vary only one acquisition parameter between the training and testing data to avoid interacting effects among parameters. Denoising performances are evaluated on patient scans, simulated phantom scans and physical phantom scans using multiple types of image quality (IQ) metrics, including mean squared error (MSE), contrast-dependent modulation transfer function (MTF), noise power spectrum (NPS) and low-contrast lesion detectability (LCD).

Results: REDCNN had larger MSE when the testing data was different from the training data in reconstruction kernel, but no significant MSE difference when varying slice thickness in the testing data. REDCNN trained with quarter-dose data had slightly worse MSE in denoising 80%-dose images than that trained with mixed-dose data. The MTF tests showed that REDCNN trained with the two reconstruction kernels and slice thicknesses yielded images of similar image resolution. However, REDCNN trained with mixed-dose data preserved the low-contrast resolution better compared to REDCNN trained with quarter-dose data. In the NPS test, it was found that REDCNN trained with smooth-kernel data could not remove high-frequency noise in the test data of sharp kernel, possibly because the lack of high-frequency noise in the smooth-kernel data limited the ability of the trained model in removing high-frequency noise. Finally, in the LCD test, REDCNN improved the lesion detectability over the original FBP images regardless of whether the training and testing data had matching reconstruction kernels.

Conclusions: REDCNN is observed to be poorly generalizable between reconstruction kernels, more robust in denoising data of different dose levels when trained with mixed-dose data, and not sensitive to slice thickness. It is known that reconstruction kernel affects the in-plane NPS shape of a CT image whereas slice thickness and dose level do not, so it is possible that the noise correlation structure described by the in-plane NPS may be used as an underlying property to predict the generalizability of this CT image denoising network.

***Index Terms*—Deep learning, CT image denoising, Generalizability performance, Image quality assessment**

Corresponding author: Rongping Zeng, rongping.zeng@fda.hhs.gov

Zeng R, Jiang L and Myers KJ are affiliated with the Center for Devices and Radiological Health, US Food and Drug Administration (FDA), Silver Spring, Maryland. Lin CY and Fessler JA are affiliated with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan. Lin Q and Skopec M were affiliated with the FDA when they were involved in this research work.

Fundings: Lin CY was supported in part by FDA Critical Path funding and Fessler JA was supported in part by NSF grant IIS 1838179.

1. INTRODUCTION

CT imaging is widely used in modern medicine for almost every disease or condition. It is highly recommended that the x-ray dose be as low as reasonable in CT exams for patient safety while maintaining the CT image quality to avoid misdiagnosis. Various approaches have been developed toward low-dose CT through improved hardware design such as automatic exposure control, kV optimization and dynamic bowtie filters [1, 2], and through advanced image reconstruction/denoising methods, such as statistical and model-based iterative reconstruction (IR) algorithms [3, 4]. Deep learning (DL) methods are now being developed for this purpose, thanks to the availability of software tools and increased computational power. Publications on applying DL in low-dose CT image denoising are growing rapidly [5-10]. Commercial DL products have become available on some CT scanners, such as AiCE from Canon Medical Systems and TrueFidelity from GE Healthcare, both receiving FDA clearance in 2019.

DL methods have been shown to be capable of improving image quality over FBP, similar to state-of-the-art iterative denoising methods [9, 11-13]. However, unlike IR algorithms that are derived based on imaging physics and data statistics, a DL method relies on training data to optimize the network coefficients to attain a noise reduction function. This data-driven mechanism makes the DL performance less generalizable when applied to processing data of different distribution from that of the training data. In most cases, characterizing the underlying data distribution to circumscribe the performance generalizability zone (i.e., the data range for which a DL network preserves its performance) is not straightforward. In CT, image resolution and noise properties are affected by image acquisition parameters such as kVp, mA, reconstruction kernel, slice thickness, pitch, etc. Therefore, it is reasonable to investigate the generalizability performance of a DL network on data of different acquisition conditions. Changes in the network's performance between two differently acquired testing datasets could indicate a potential data distribution shift caused by the associated acquisition parameters. Thus, an analysis of the data properties associated with the acquisition parameters may provide insight on possible ways to characterize the data distributions for the generalizable range of a DL-based CT image denoising network.

Following this reasoning, we investigated a residual encoder-decoder convolutional neural network (REDCNN) for low-dose CT image denoising [5] and used patient scans from the Low Dose Grand Challenge (LDGC) dataset [14] to train that network [15]. We examined the denoising performance changes under three scenarios. In each scenario only one acquisition parameter changed between the training and testing data. The three acquisition parameters were reconstruction kernel, slice thickness and dose level. The image quality (IQ) metrics for evaluating the denoising performance included 1) mean squared error (MSE), a global IQ metric; 2) contrast-dependent modular transfer function (MTF) and noise power spectrum (NPS), standard CT IQ metrics that characterize the image resolution and noise properties; and 3) low-contrast lesion detectability (LCD), a more clinically relevant task-based IQ metric. We included these multiple IQ metrics to examine how well they support the evaluation of a denoising method's impact on task-based image quality. While a denoising algorithm may appear to beautify an image, there is the possibility that it impairs the detection or characterization of subtle signals and other image features

We compared the performance behavior of the DL denoising network under different training and testing conditions, particularly, varying the reconstruction kernel, slice thickness and dose level bewteen the training and testing data As

being mentioned previously, these acquisition parameters affect the image resolution and noise property of a CT image set. For example, reconstruction kernel changes the in-plane resolution and noise correlation structure. Slice thickness mainly affects the z-direction resolution. Dose level determined the noise magnitude. A degradation in the DL network's denoising efficiency due to a mismatch of a data acquisition parameter would be associated with the underlying data properties that are altered by that acquisition parameter. Based on the findings regarding whether a change in each of the three parameters causes a substantial degradation in the DL's denoising perfomrances or not, we may learn and identify which underlying data properties are most important in predicting the denoising network's generalizability.

The rest of the paper is organized as follows. Section 2 explains the low-dose CT denoising network, the training scheme for preparing the generalizability tests, the evaluation methods and testing data. Section 3 presents the results. Section 4 discusses our observations on the DL generalizability performance followed by the conclusions.

2. METHODS

2.1. *Low-dose CT denoising network*

Let $\boldsymbol{x} \in R^{m\times n}$ denote a low-dose CT reconstructed image; the DL-based denoising problem is to optimize the network $C(\boldsymbol{x})$: $R^{m\times n} \rightarrow R^{m\times n}$ that maps $\boldsymbol{x}$ to its corresponding high-dose image $\boldsymbol{y} \in R^{m\times n}$ by minimizing a loss function between $\boldsymbol{x}$ and $\boldsymbol{y}$ over a given set of training data. After the network is optimized, a noisy CT image can be passed through the network to produce an image intended to have reduced noise.

Various network structures have been explored in the literature for low-dose CT image denoising. Some typical networks include convolutional neural networks [6], residual networks [5, 10, 16, 17], UNet [8, 18] and Generative adversarial networks [7, 19]. For this paper, we selected the residual encoder-decoder convolutional neural network (REDCNN) developed by Chen et al.[5] as a denoising example for the generalizability test. Our emphasis here is not on the demonstration of an innovative denoising algorithm, but rather the illustration of an approach for assessing DL generalizability. We come back to this point in the discussion.

As illustrated in Fig.1, REDCNN contains ten layers, the first five being convolutional layers and the last five being deconvolutional layers. A rectified linear unit (ReLU) activation function follows the convolutional or deconvolutional operator in each layer. Residual learning is realized by including three shortcuts connecting the convolution layer and deconvolution layer. All the convolutional and deconvolutional layers have a filter size of 5×5. The number of filters is 96 for all the layers except that the last layer has one filter. For more details about the network design, please refer to [5]. We selected this residual network design because it was not very complicated but has been shown to have potential for effective CT image denoising similar to some traditional iterative denoising methods under the conditions tested in the papers by Chen et. al. and Zeng et. al. [5, 20].

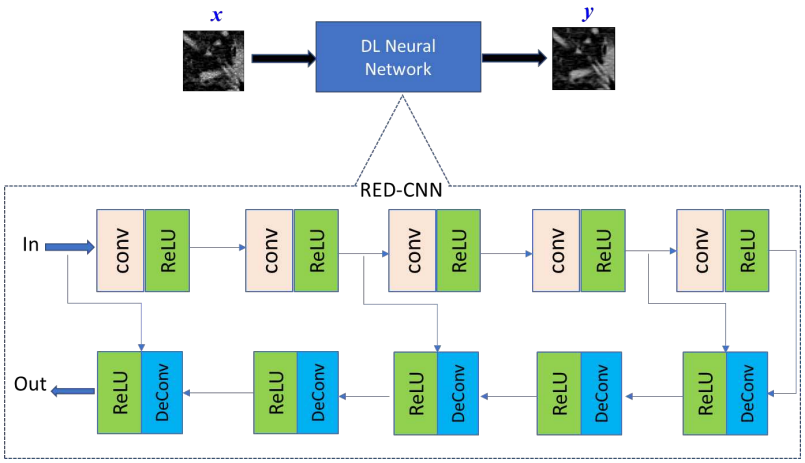


Fig. 1. Illustration of the REDCNN denoising network.

The loss function for training the denoising network we used was the MSE between the network output and the corresponding high-dose target images. Some investigators add terms to the loss function to encourage image smoothness and feature similarity, or to regularize the network parameters with weight decay to avoid overfitting [8, 10]. However, we focused on the most commonly used MSE loss function in this work.

2.2. *Training data categorization*

The denoising network was trained using the patient scans in the Low-Dose Grand Challenge (LDGC) dataset [14]. There are ten datasets in LDGC covering chest to abdomen. Each patient dataset contains a full-dose scan acquired on a Siemens Somatom Definition AS+ or Definition Flash scanner model and a simulated quarter-dose scan. Each scan was reconstructed with two slice thicknesses (1 mm and 3 mm) and two reconstruction kernels (a sharp kernel named D45 and a smooth kernel named B30). The corresponding quarter- and full-dose image pairs were treated as training input and training target in the DL training process, respectively. Among the ten patient datasets, seven patient datasets were used for training since more data were needed to train than test the network that contained more than 1.8 million coefficients. 350 slices of size 512x512 were randomly selected from the seven patients and each slice was divided into 55x55 patches excluding the air patches outside of the body, resulting in about 70,000 training patches in total.

The variety of reconstruction thickness, reconstruction kernel and dose level make the LDGC datasets suitable for this performance generalizability study. We grouped the CT volumes into three pairs of training data according to the acquisition parameters as shown below. In each pair, only one acquisition parameter value was varied to avoid interacting effects among the parameters.

Dose level effect:
- Smooth kernel / 3 mm thickness / **25% dose level**
- Smooth kernel / 3 mm thicknesS / **Mixed dose levels**

Kernel effect:
- **Sharp kernel** / 3 mm thickness / mixed dose level
- **Smooth kernel** / 3 mm thickness / mixed dose level

Thickness effect:
- Smooth kernel / **1 mm thickness** / mixed dose level
- Smooth kernel / **3 mm thickness** / mixed dose level

With this data arrangement, we can obtain three pairs of trained DL networks. For convenience, we name the networks according to the parameter setting of the training data as follows: *DLkernel-thickness-dose*. For example, "DLsharp-

3mm-mix%" represents the REDCNN trained with images of sharp kernel, 3mm thickness and mixed dose levels; "DLsmooth-1mm-25%" represents the REDCNN trained with images of smooth kernel, 1mm thickness and a single 25% dose level. Each pair of networks was cross evaluated on two types of test sets to determine how the performance may change when the testing data were acquired with a different parameter value.

There was only one reduced dose level (25%) available in LDGC. The mixed-dose data were synthesized using the full and quarter-dose scans by a simple blending of the two scans: A noise map was obtained by subtracting the quarter-dose image from the full-dose image and then a portion of the noise map was blended back into the full-dose image as follows:

$$\boldsymbol{x}_d = \boldsymbol{x}_f + \alpha\,(\boldsymbol{x}_q - \boldsymbol{x}_f),\ \alpha \geq 0,$$

where $\boldsymbol{x}_d, \boldsymbol{x}_f,$ and $\boldsymbol{x}_q$ represent the synthesized noisy image at a dose level d, the original full-dose and the quarter-dose images, respectively. The scaler α denotes the blending factor. When $\alpha = 1$, the outcome is exactly the quarter-dose image. When $\alpha = 0$, the outcome is the full-dose image. For an arbitrary nonnegative α, the outcome corresponds to $1/((1-\alpha)^2 + 4\alpha^2)$ of the full-dose scan. We varied the blending factor randomly in the interval of [0.5, 1.2] for the mixed dose training data case, resulting in images of dose levels ranging from 17% to 80% of the full-dose level.

2.3. Performance evaluation

To evaluate the performance, we considered the following IQ metrics: MSE, contrast-dependent MTF, NPS, and LCD. MSE reflects how well the network performs in minimizing the loss function that it is designed to do. We did not evalute the other global metrics like PSNR or SSIM in this work since they are highly correlated with MSE. However, it is well known that a denoised image with smaller MSE does not necessarily have better diagnostic image quality. We included the standard CT IQ metrics MTF and NPS as they are commonly used to characterize the image resolution and noise texture. Lastly, we evaluated the denoising performance in terms of LCD, a task-based IQ metric measuring the capability of detecting low-contrast lesions in the denoised images.

2.3.1. Mean Squared Errors (MSE) test

For the MSE measure, the slices from one patient dataset in LDGC that were not included in the training were used as a test set. The total slice numbers were more than 200 slices and 500 slices for the testing cases of 3mm and 1mm slice thickness CT volumes respectively. For each slice, the full-dose image was used as a reference to calculate the MSE $\left(= \frac{\|\text{Noisy Image} - \text{Ref Image}\|^2}{\text{The total number of pixels}}\right)$ before and after the DL denoising. Then the MSE reduction rate $\left(= \frac{\text{MSE before denoising} - \text{MSE after denoising}}{\text{MSE before denoising}} \times 100\%\right)$ was calculated to quantify the denoising performance. Based on the multiple slices in the test CT volumes, statistics of the MSE reduction rates can be obtained and compared between the pairs of DL networks.

2.3.2. Contrast-dependent Modular Transfer Function (MTF) and Noise Power Spectrum (NPS) test

We simulated 2D phantom CT scans for the MTF and NPS tests. We also collected multi-slice CT phantom scans to validate the simulation-based results, which are described in Sect 2.4. For the MTF measure, a contrast phantom (Fig. 2) similar to the CATPHAN600 contrast module was simulated. The contrast phantom contained eight disks of 2 cm

diameter, similar to the HU contrasts contained in the CATPHAN600 phantom. Contrast-dependent MTF was measured using the methods described in [21]. Note that a noiseless CT scan of the contrast phantom was simulated for the MTF test to eliminate any uncertainties caused by random noise, since MTF represents a deterministic behavior of an imaging system. For the NPS measurement, 50 noisy water phantom CT scans were simulated. A Region of Interest (ROI) of size 64 x 64 pixels at the image center was extracted from each realization. Local NPS was estimated by taking the average of the modulus square of the Fourier transform of the noise images after being subtracted from the mean of the 50 realizations.

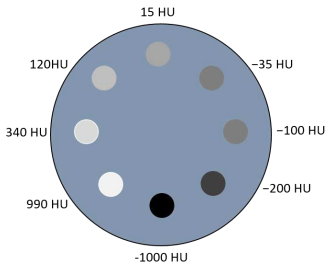


Fig. 2: Sketch of the digital contrast phantom that is used to measure the MTF. It mimics the CATPHAN600 contrast module with an added disk of 15HU contrast.

The simulated CT scans were created from a virtual fan-beam 2D CT scanner. The virtual scanner had distances of 595 mm from the x-ray tube to the isocenter and 1085.6 mm to the detector, the same as those in the Siemens CT scanner used to collect the LDGC dataset. Poisson noise was modeled at the detector but electronic noise was not. We varied the air photon flux to achieve different noise levels. To simulate the reconstruction kernels in the LDGC data, two Hann filters of different cutoff frequencies (named Hann1 & Hann2) were used in our FBP reconstruction. The cutoff frequencies were tuned to closely match the MTF50% and MTF10% of the D45 and B30 filters (see Table 1). Note that MTF50% and MTF10% are the frequency values where MTF drops to half and 10% respectively. For convenience, we refer to Hann1 and D45 as sharp kernels, and Hann2 and B30 as smooth kernels in this paper. The reconstruction pixel size was set to 0.664 mm, corresponding to a 512×512 reconstruction matrix of a 340 mm field of view (FOV). Since we only simulated 2D scans, slice thickness was not a modeled parameter in the virtual scanner. The simulated scans could be treated as a very thin slice thickness setting. The CT simulation code was implemented based on the Michigan Image Reconstruction Toolbox (MIRT) that is available online at https://web.eecs.umich.edu/~fessler/code.

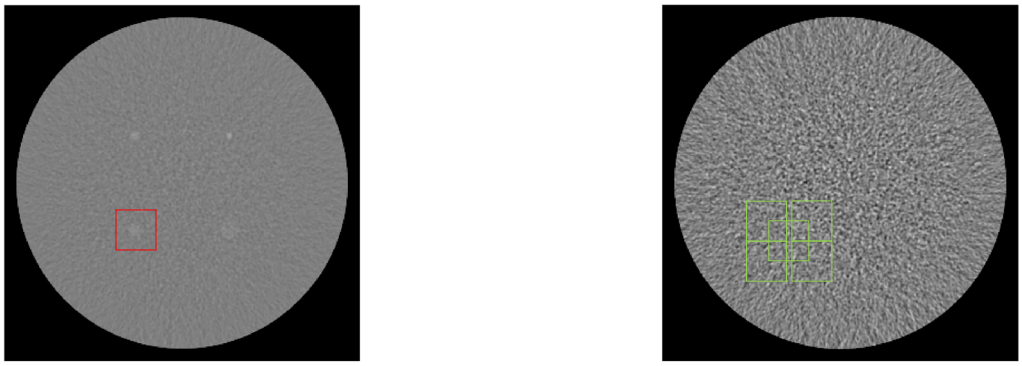
Table 1. The MTF50% and MTF10% values in lp/cm of the commercial reconstruction kernels (D45, B30) and simulated reconstruction kernels (Hann1 and Hann2).

Resolution (lp/cm)	D45 (sharp)	Hann1 (sharp)	B30 (smooth)	Hann2 (smooth)
MTF50%	5.6	5.6	3.5	3.5
MTF10%	9.4	10.4	5.9	6.2

2.3.3.Low-Contrast Detectability (LCD) test

Low-contrast detectabilites were estimated using a model observer and simulated phantom images containing very low-contrast objects. Specifically, we simulated 200 CT scans of the signal module and 100 scans of the background module of the MITA-LCD phantom CCT189 (Fig. 3) at five exposure levels. The signal module contained four low-contrast disks with varying size/HU combinations (3mm/14HU, 5mm/7HU, 7mm/5HU, 10mm/3HU) to mimic sutble

lesions. Five exposure levels were simulated: 100%, 85%, 70%, 55% and 30%. The 100% dose level corresponded to an air photon count of 3×10^5 per detector pixel. For each disk signal, a signal-present (SP) ROI was cropped from the scan of the signal module and five signal-absent (SA) ROIs were cropped from the background module at the vicinity of the signal location. A Laguerre-Gauss channelized Hoteling model observer (LG-CHO) was applied to estimate the signal detectability [22]. The LG-CHO had five channels and the Gaussian width was adjusted to match the size of the disk to be detected. Among the 200 SP ROIs and 500 SA ROIs, 80 pairs of SP and SA ROIs were used to train the model observer. The remaining ROIs were used to estimate the detectability, quantified by the area under the receiver operating curve (AUC).



(a) (b)

Fig. 3: Sample CT images of the signal module (left) and the background module (right) for the LCD test. Red and green boxes illustrate the locations for cropping signal-present ROIs and the corresponding signal-absent ROIs. Note that the CT image of the signal module shown here is an average of 20 realizations from the highest dose level reconstructed with FBP of smooth kernel to make the low-contrast signals visible. The display window is [-50 50] for both images.

2.4. *Validation with physical phantom scans*

CT scans of a CATPHAN600 phantom (The Phantom Laboratory, Salem, NY) were collected on a Somatom Definition AS model (Siemens Medical Solutions USA, Inc, Malvern, PA) to validate the observations in the MTF and NPS test with simulated phantom scans. The scan protocols were designed to closely match the settings in the LDGC dataset, including the parameters of kVp, x-ray filter, detector collimation, slice thickness, convolution kernel and reconstruction field of view. Table 2 provides a summary of those major scan parameters in the LDGC, together with the parameter settings for our phantom scans. As can be seen from the table, the reconstruction kernel and the slice thickness were the same for the LDGC patient scans and the phantom scans. However, there existed some differences in the other parameters as discussed next.

First, we turned the automatic exposure control (AEC) off since "on" or "off" would not matter much for a cylindrical phantom with minor interior background variation. The patient scans had kVp varying in the range of 100 – 120 kV across the slices due to AEC. For our phantom scan, the kVp was fixed at 120 kV. Second, we scanned the phantom with three dose options, named high-dose, full-dose and quarter-dose. The full-dose option was set to match the average values of the CTDI of the full-dose patient scans. The high-dose option (higher than the full-dose option) was added to reduce the uncertainty in the MTF estimations. Third, for the x-ray filter setting that may affect the x-ray spectrum shape, we used "FLAT" filter since most of the patient scans were with this option. Fourth, our phantom scans had the

same single collimator width 0.6 mm as the LDGC patient scans. However, the total collimator width was 12 mm, narrower than 38.4 mm in the LDGC scans, because the 38.4 mm collimator option was not available on the scanner model we used. Fifth, the pitch factors in the patient scans varied from 0.6 to 0.8. In our phantom scan, the pitch was set to 0.8 to save scan time. As long as the pitch factor was smaller than 1, degradation in the z-directional sampling would be negligible for the scans of the cylinder-shaped CATPHAN600 phantom. Lastly, the reconstruction field of view (FOV) varied in the patient scans, ranging from 340 to 420 mm due to the different patient sizes. Reconstruction FOV affects the pixel size. For the phantom scans, we set the FOV to be 380mm, close to the average FOV of the 10 patient scans. This resulted in a pixel size of 0.74 mm in the reconstructed phantom volume.

In total, we collected one high-dose scan, and five repeats of the full-dose and quarter-dose scans. For each scan, reconstructions with 1 mm and 3mm slice thickness, sharp and smooth kernel were generated, resulting in 44 CT volumes.

Table 2: Comparison of the data acquisition parameters between the LDGC dataset and our phantom scans.

Dataset	AEC	kVp (kV)	CTDI (mGy)	x-ray filter	Single/Total Collimator width (mm)	Pitch	FOV (mm)	Slice Thickness (mm)	Reconstruction kernels
LDGC	XYZ-EC	100 - 120	19.7 (mean for Full)	FLAT (8) WEDGE_3 (2)	0.6 / 38.4	0.6 to 0.8	378 (mean)	3 1	B30f D45f
Phantom scans	OFF	120	32.1 (High) 20.0 (Full) 5.0(Quarter)	FLAT	0.6 / 12	0.8	380	3 1	B30f D45f

3. RESULTS

3.1. *Mean Squared Errors*

Fig. 4 shows box plots comparing the MSE reduction rates of the three pairs of DL networks. For the dose effect (Fig. 4a), when tested on the quarter-dose images, the DL networks trained solely with quarter-dose data and trained with mixed-dose data had almost equivalent MSE reduction rate. When tested on the 80% dose images, the DL network trained with mixed dose reduced MSE noticeably more. This indicates that the DL denoising network trained with mixed-dose data generalized better on data of different dose levels. For the reconstruction kernel effect, Fig. 4b shows that when the training and testing data had a different reconstruction kernel, the DL network performed subtantially worse. This indicates that the DL denoising network did not generalize well on data with a different reconstruction kernel. For the thickness effect (Fig. 4c), in both the 3mm and 1mm thickness testing cases, the MSE reduction rate was similar between the DL networks trained with the two different thickness datasets. The DL network trained with 3mm thickness appeared to be slightly better at maintaining testing performance across thicknesses, but the difference was not statistically significant since the two distribution ranges heavily overlapped. The similar performances indicate that the slice thickness parameter may not be critical to the DL denoising network.

Fig. 5 presents sample CT images to visually demonstrate the effect of reconstruction kernel. As can be seen , in the test case of FBP smooth (top two rows in Fig. 5), the DLsharp-3mm-mix% processed image obviously appears to be much noiser than the image processed by DLsmooth-3mm-mix%. Meanwhile, in the test case of FBP sharp (bottom two rows in Fig. 5), the image texture of the DLsmooth-3mm-mix% processed FBP sharp image appears quite different from the others. It is also noticeable that the anatomical structures in the DLsmooth-3mm-mix% processed image slice are oversmoothed and some small features are lost.

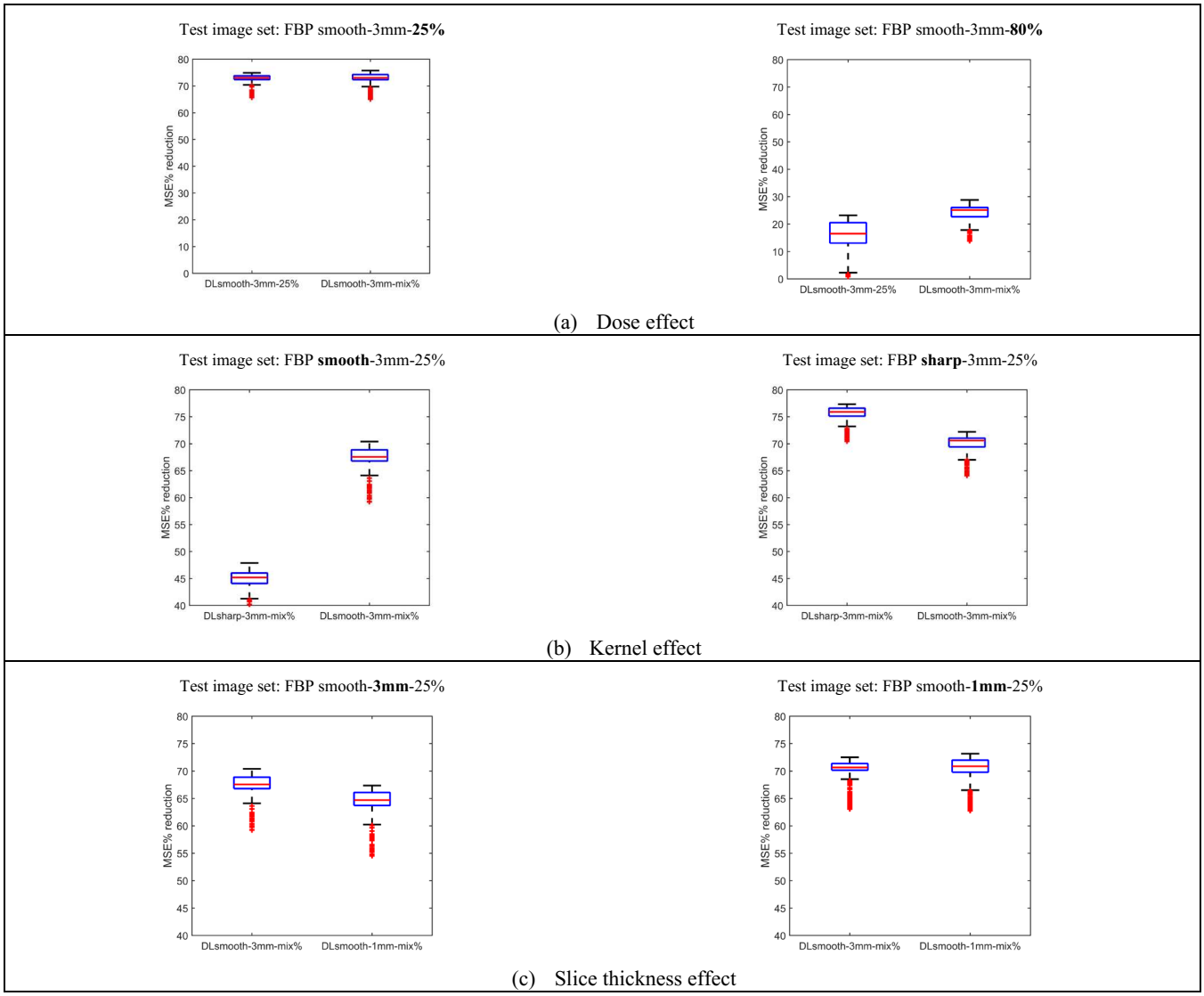
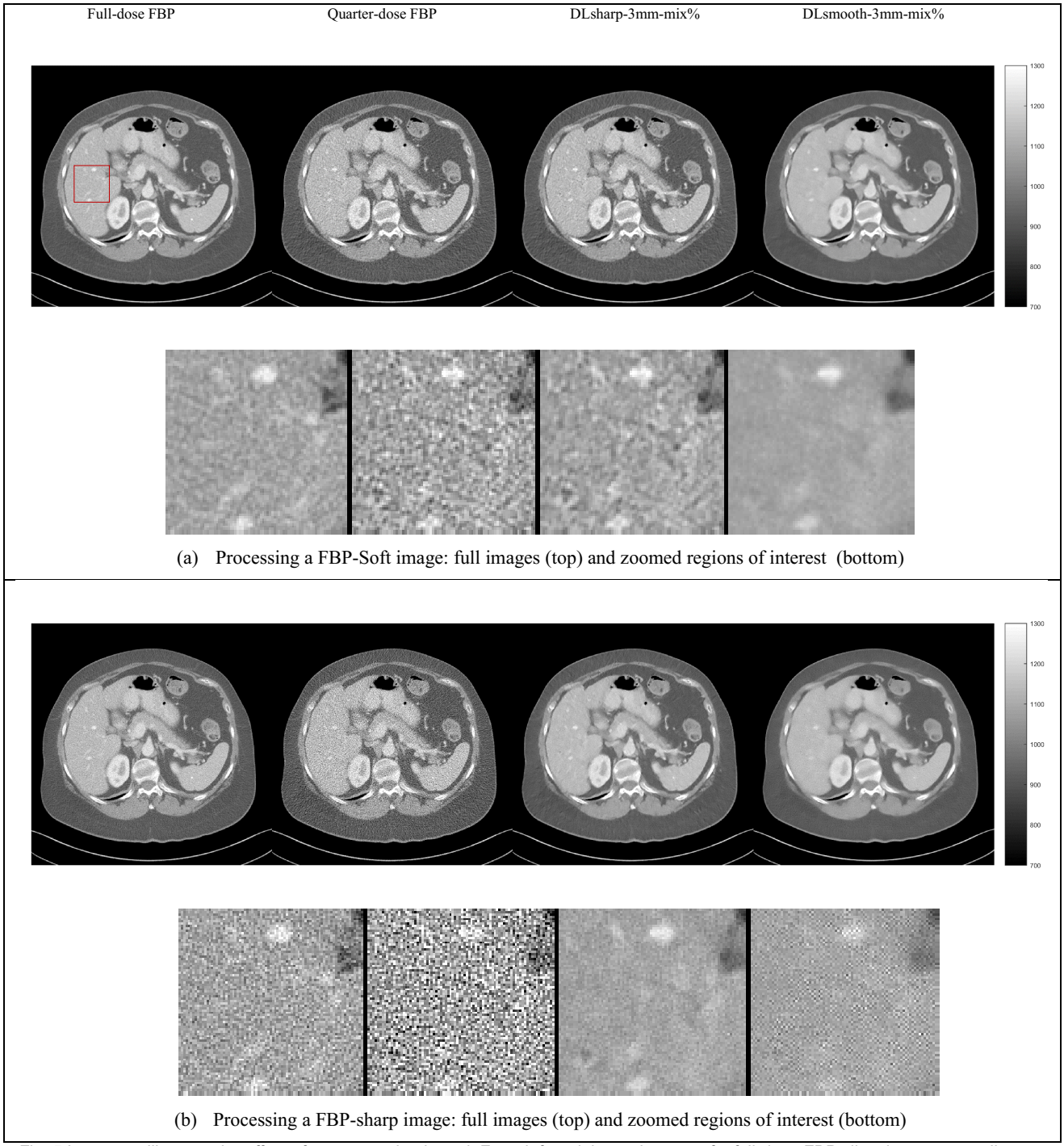


Fig. 4 Effects of the training parameters on the MSE reduction rate of the DL networks. The first row compares the dose level effect, the second row compares the reconstruction kernel effect and the third row compares the thickness effect.



(a) Processing a FBP-Soft image: full images (top) and zoomed regions of interest (bottom)

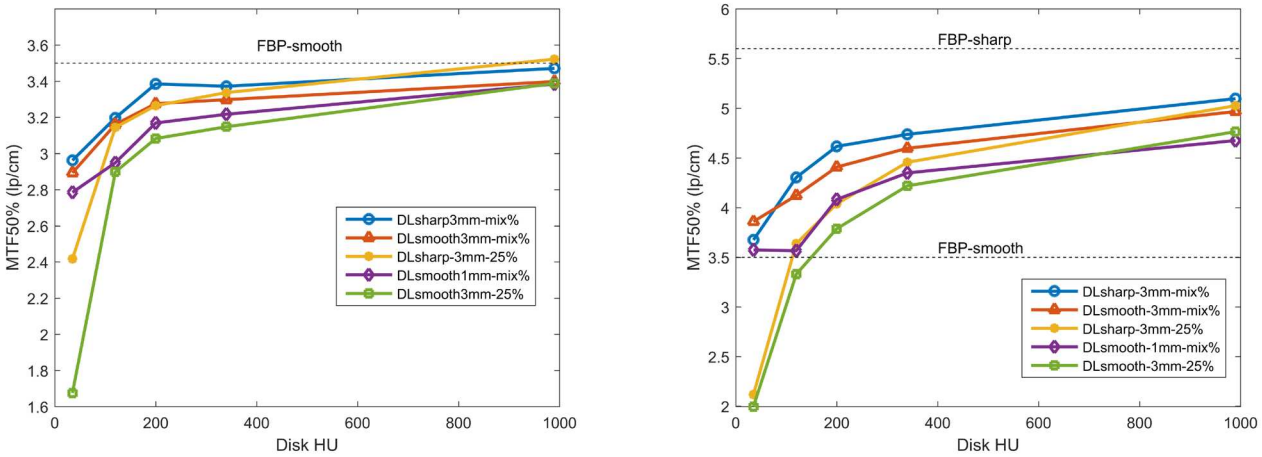
(b) Processing a FBP-sharp image: full images (top) and zoomed regions of interest (bottom)

Fig. 5 Images to illustrate the effect of reconstruction kernel. From left to right are images of a full-dose FBP slice, its corresponding quarter-dose FBP slice, DLsharp-3mm-mix% and DLsmooth-3mm-mix% denoised quarter-dose slice. a). for processing a quarter-dose FBP image reconstructed with smooth kernel and b) for processing a quarter-dose FBP image reconstructed with sharp kernel. The red box in the full-dose FBP image in (a) indicates the region of interest that is zoomed for display.

Contrast-dependent Modular Transfer Function

In this test, we generated a noiseless sinogram of the contrast phantom and reconstructed the sinogram with FBP using sharp and smooth kernels. The noiseless FBP images were then processed by the DL networks. Contrast-dependent MTF curves were estimated at these five contrasts: 990, 340, 200, 120 and 35 HU. The MTF50% value was calculated for each MTF curve and plotted as a function of the HU contrast to characterize the contrast-dependent image resolution.

Fig. 6a and 6b show the contrast-dependent image resolution curves for the DL networks in processing FBP-smooth and FBP-sharp images, respectively. The curves clearly show that the image resolution decreases with contrast. This nonlinear smoothing behavior is similar to that of traditional iterative reconstruction and denoising methods. We also see that the curves in Fig 6a and 6b show the same contrast-dependent trends for both smooth and sharp FBP input recontructions: the DL network trained with sharp-kernel data had slightly better image resolution (higher MTF50% value) than the DL network trained on smooth-kernel data; the DL network trained with thicker slice data had slightly better image resolution than the DL network trained with thinner slice data; the DL network trained with mixed-dose data had slightly better image resolution than the DL network trained with single-dose data, except at the contrast level of 35 HU where the resolution dropped greatly for the single-dose DL network. In summary, the trends in the MTF test indicate that the image resolution of the DL denoising network was not very sensitive to the kernel and slice thickness parameters. However, it appears that with mixed-dose training data, low contrast resolution was much better preserved.



(a) (b)

Fig. 6: Contrast-dependent MTF50% curves of the DL networks for processing the FBP-smooth images (a) and FBP-sharp images (b).

3.3. Noise Power Spectrum

We simulated 50 noisy scans of a cylindrical water phantom for the NPS estimation, with an air photon count of 2.4×10^5 per pixel. Each noisy scan was reconstructed by FBP for both sharp and smooth kernels. Then the noisy images were processed by the DLsharp-3mm-mix% and DLsmooth-3mm-25% to compare the effect of kernel in the NPS test. Note that we did not further examine the effects of the slice thickness and dose level parameters in the NPS and the LCD test, because the pervious MSE and MTF test results showed that the DL network trained with thicker slice thickness and mixed-dose data had better performances. For convenience, we simplify the names of DLsharp-3mm-mix% and DLsmooth-3mm-mix% as *DLsharp* and *DLsmooth* afterward in the NPS and LCD test.

Fig. 7 presents the NPS images and Fig. 8 plots the corresponding radial profiles. The radial profiles clearly show that the DL networks reduced the noise magnitude and shifted the peak frequency toward zero. Again, this is a behavior similar to that of traditional iterative reconstruction and denoising methods. In general, DL denoised images had noise components concentrated more in the lower frequency bands compared to the original FBP images. However, one may

notice a contrasting appearance in the NPS of DLsmooth processed FBP-sharp image (the rightmost in Fig 7b): much higher magnitude at the four corners (high-frequency regions). The 1D radial profile clearly shows that the corresponding NPS curve has a rising tail (as indicated by the arrow in the Fig. 8b) after about 5 lp/mm. Moreover, the tail's shape and height closely match those of the NPS curve of the original FBP-sharp images, indicating that the high-frequency noise was not removed by the DLsmooth network. An example CT image patch from a uniform background (Fig. 8c) also demonstrates the remaining high-frequency noise in the DLsmooth processed FBP-sharp images, appearing as tiny checker-board like artifacts. This phenomenon suggests that the DLsmooth network possibly did not learn to remove the high-frequency noise from the smooth kernel training data, since the training data did not contain noise in the high-frequency band.

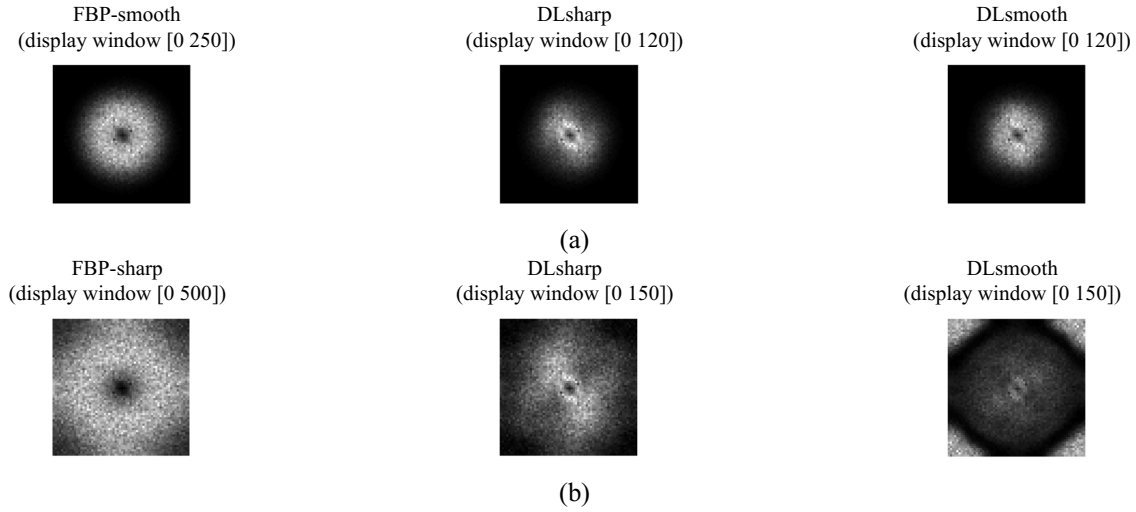


Fig. 7: 2D NPS of the original FBP images and the corresponding DLsharp and DLsmooth processed images. Results on FBP smooth kernel is in (a) and on FBP sharp kernel is in (b).

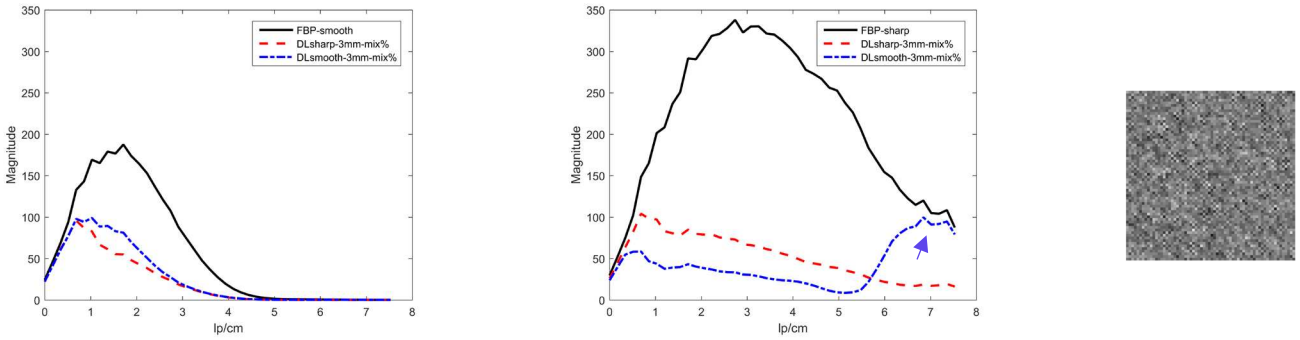


Fig. 8: The 1D radial profiles of the NPS of the original images and the corresponding DLsharp and DLsmooth processed images in (a). The 1D NPS radial profiles of the original FBP-sharp images and the corresponding DLsharp and DLsmooth processed images in (b). The blue arrow in (b) indicates the raised tail in the NPS of the DLsmooth processed FBP-sharp images and the sample image patch in (c) illustrates the remaining high-frequency noise, which appears as tiny checker-board like artifacts.

3.4. MTF and NPS test using physical phantom CT scans

We conducted the MTF and NPS tests again using the physical CT scans of CATPHAN600 to validate the observations found in the results using simulated phantom CT scans.

First, we measured the contrast-dependent image resolution of the DL networks processing 3mm thickness high-dose FBP images. Fig. 9 displays the resolution curves. Due to image noise, the MTF function estimated from the disks of contrast below 100HU were not reliable. Therefore, the contrast-dependent image resolution curves were based on the disks of air, PMP, LDPE and Polystyrene in the CATPHAN600 contrast module, which had measured mean absolute contrast of 1100, 260, 170 and 115. The resolution curves in Fig. 9 also show that DL networks trained with data of sharp kernel, thicker slice thickness, mixed-dose levels had better image resolution than their counter parts, similar to the findings obtained in the testing results with simulated 2D CT scans.

Second, we estimated the NPS images and extracted their 1D radial profiles of the DL networks processing 3mm thickness full-dose FBP images, as shown in Fig. 10. A rising tail in the NPS curve of the DLsmooth processed FBP-sharp images was also observed, similar to that in Fig. 7b. We omitted the NPS results for processing the low-dose FBP images since they present similar trends. These experiments showed that the NPS results obtained with the physical phantom CT scans agreed with those obtained with the simulated CT scans.

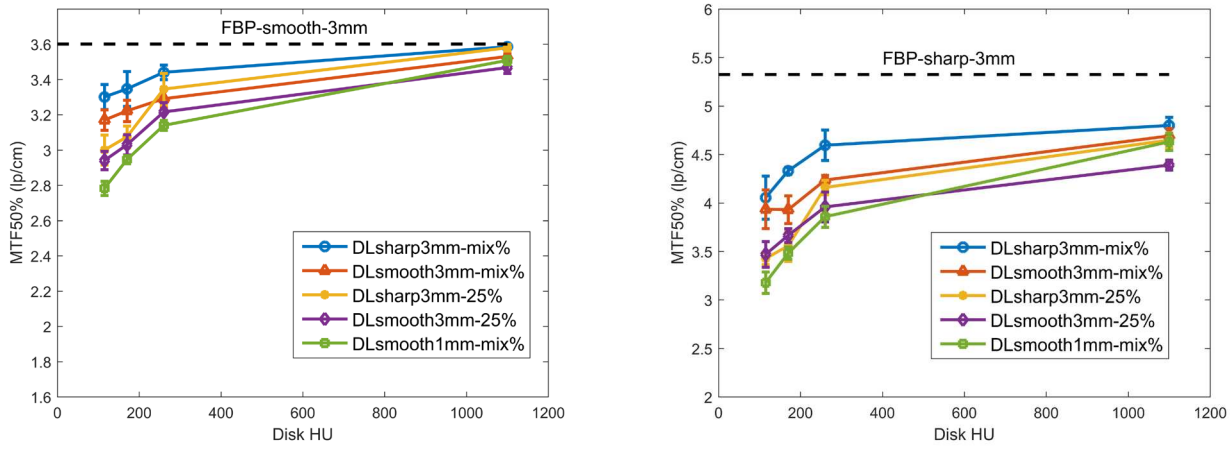


Fig. 9: Contrast-dependent MTF50% curves of the DL networks for processing the FBP-smooth-3mm images in (a) and FBP-sharp-3mm images in (b) using the physical phantom CT scans.

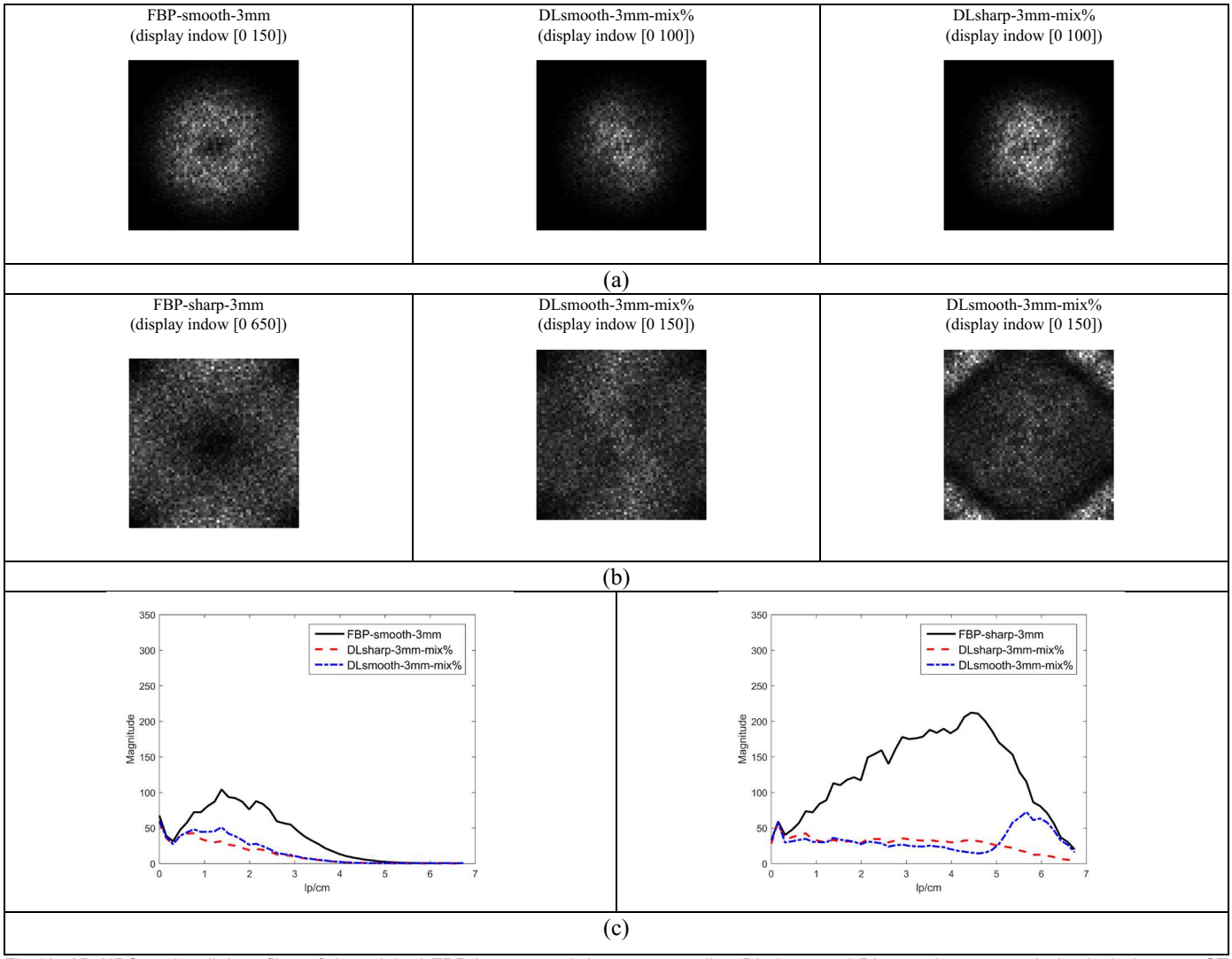


Fig.10: 2D NPS and radial profiles of the original FBP images and the corresponding DLsharp and DLsmooth processed physical phantom CT images. Results for processing FBP smooth images are in (a) and the left plot in (c). Results for processingFBP sharp kernel are in (b) and the right plot in (c).

3.5. *Low Contrast Detectability*

Fig. 11 plots AUC, a measure of low-contrast detectability, as a function of dose for detecting the 10mm/3HU inserts in the simulated MITA-LCD phantom. As can be seen in the figure, both the DLsharp and DLsmooth networks improved the detectability over the original FBP images regardless of the original reconstruction kernels. The DLsmooth network had similar AUCs as the DLsharp network in processing FBP-smooth images but significantly higher AUCs in processing FBP-sharp images. We will explain the possible reasons later in the discussion. The detectability curves are not shown here for the other three inserts (3mm/14HU, 5mm/7HU, 7mm/5HU). In general, we observed that the detectability curves in the original FBP images and the DL denoising images were almost the same for detecting the two smaller inserts (3mm/14HU and 5mm/7HU), then became more separated as the size of the insert increased, but the relative performance trends were the same for detecting these inserts. Therefore, we only present the curves for detecting the 10mm/3HU insert since the curves separated the most in this case.

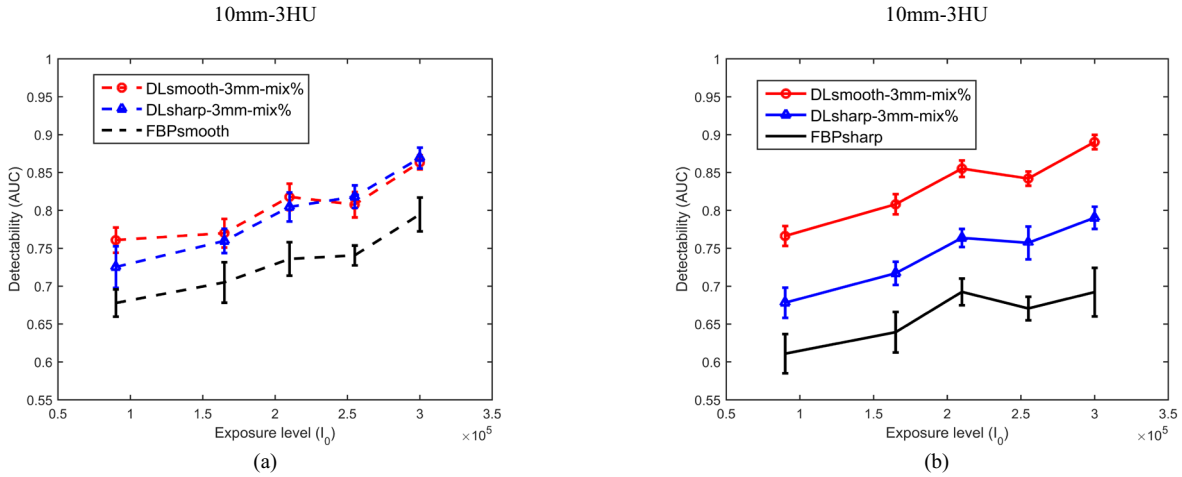


Fig. 11 Detectability curves for the insert (10mm-3HU) in the original FBP images, and denoised FBP images with DLsharp and DLsmooth: (a) for processing FBP sharp images and (b) for processing FBP smooth images

4. DISCUSSION

In this work we presented a framework for the evaluation of performance generalizability of a DL-based CT image denoising method, using the REDCNN as an example denoising algorithm. We used the patient CT scans in the LDGC dataset to train the network on data acquired with different acquisition parameters. Based on the data variety, we examined the performance generalizability of the denoising network on three parameters: reconstruction kernel, slice thickness and dose levels. Performances were evaluated using MSE, contrast-dependent MTF, NPS and LCD. We observed the following three points from the testing results.

First, the denoising network did not generalize well between the sharp and smooth reconstruction kernels. This is reasonable since the reconstruction kernel is the most dominant factor that determines the noise correlation structure in a FBP reconstructed image. The NPS curves of the FBP-sharp and FBP-smooth images in Fig. 8 & 10 obviously differ in both the peak and the cutoff frequencies. Due to the DL's data-driven mechanism, a denoising network may not recognize noise components that are not seen in its training data. This explains the remaining high-frequency noise in the DLsmooth processed FBP-sharp images. On the other hand, the image resolution property was not much different between the DLsmooth and DLsharp networks since the denoising network was not trained to alter image resolution.

Second, the denoising network was not sensitive to slice thickness. When all the other acquisition parameters are kept the same, a 3mm slice thickness CT volume may be considered as being formed by a moving average (or weighted average) of every three adjacent slices of the 1mm slice thickness CT volumes. Averaging along the longitudinal direction does not alter the noise correlation structure within a slice, so the denoising networks trained with 3mm and 1mm thickness image slices were not much different. However, the noise magnitude in a 3mm thickness slice is usually lower than that in the corresponding 1mm slice. In this sense, the target images in the 3mm thickness training data had slightly better image quality, which may explain why the DL-3mm network performed slightly better than the DL-1mm network in both the MSE and MTF tests.

Third, the denoising network was more robust in processing images of an unknown noise level when trained with mixed-dose data. The MSE results showed that the DL-mix% network maintained the MSE reduction rate in processing quarter-dose slices and reduced MSE more when processing slices of a different dose level than the DL-25% network.

The DL-mix% also preserved the low-contrast image resolution better, as shown in the MTF test where the testing data may be considered as a very high-dose scan. Since the noise correlation structure did not change except the magnitude in the various dose level settings, training with mixed-dose data increased the adaptivity of the network in processing CT images with unknown noise levels. The finding on the dose parameter agrees with the observation in Chen et. al. [6], where a three-layer convolutional neural network (CNN-3) trained with mixed-dose data was found to have better denoising performance than the CNN3 trained with single-dose data in processing data at all the tested noise levels. Mixing the data of different dose levels in training can also be considered as a data augmentation strategy that is commonly used to improve robustness of a DL network performance [23, 24].

Despite the finding based on the MSE and NPS tests that the denoising network did not generalize well between reconstruction kernels, the DLsmooth network surprisingly achieved much better detection performance than the DLsharp network in detecting the 7mm and 10mm disks after processing the FBP-sharp images. It appears that the remaining high-frequency noise in the DLsmooth processed FBP-sharp images did not negatively affect these detection performances. The reason could be that the signal information of the four disks mostly concentrated in the lower frequency band such that the high-frequency information was not used by the model observer in the detection tasks. As shown in Fig. 7b, the rising tail of the NPS curve of DLsmooth starts at about 5 lp/cm. Even for the smallest 3mm disk, its main spectrum lobe is within 3.3 lp/cm; the signal power of most of the low-contrast disks included in the LCD phantom already dimishes at 5 lp/mm. Based on the MTF and NPS tests, the DLsmooth appeared to have comparable resolution and better noise reduction in the lower frequency band compared to DLsharp, which may have contributed to the higher detectabilities of DLsmooth in the LCD test. The results and our analysis indicate the limitation of this LCD test in evaluating the overall performance of DL denoising networks. Additional tasks focusing on high-frequency information need to be developed to allow a thorough evaluation of a DL method's denoising performance, such as shape discrimination, size estimation, etc.

Due to the limited data variety in LDGC, we examined the performance generalizability only on three CT acquisition parameters in this work. Other parameters associated with a CT scan can also affect the FBP image quality, such as kV, helical pitch, detector collimation width and scan FOV. It is worth discussing how the DL denoising network REDCNN may generalize across other parameters. As is known, a DL network usually generalizes well within its training data distribution. In a FBP-reconstructed CT image, the noise approximately follows a correlated multi-variate Gaussian distribution. The noise correlation structure can be described by the (local) NPS. The results in this study provide evidences to support the hypothesis that NPS may be used as an underlying property to predict the generalizability performance for REDCNN denoising algorithm among different CT acquisition parameters: if a different parameter value associated with the testing data does not alter the NPS shape relative to the training data, the DL network will maintain its denoising performance, such as between the two different thickness settings and between different dose levels; If a different parameter value substantially changes the NPS shape, the DL network will likely have poorer denoising performance, such as between the sharp and smooth reconstruction kernels. Based on this hypothesis, we make the following predictions on the generalizability related to other scan parameters.

Since the kV setting mainly affects the image contrast and not the noise color, we expect a denoising network to generalize well in the typical kV range (80-140 kVp) of CT scans. Helical pitch and detector collimation width mainly affect the longitudinal resolution, similar to the effect of the slice thickness parameter. Therefore, the denoising network should not be sensitive to the change of these two parameters as well. The scan FOV (or reconstruction FOV) setting usually varies with the patient size. With a fixed CT reconstruction matrix size (512×512), the scan FOV setting determines the pixel size of the reconstruction grid, i.e., the image-domain sampling frequency. Backprojecting the noisy sinogram to a finer or a coarser image grid will affect the noise correlation between adjacent image pixels. Therefore, the NPS of CT scans reconstructed with different FOVs will be different. If the FOV setting changes significantly, such as from average-size patients to obese patients or to pediatric patients, the denoising performance may not generalize well. We will conduct experiments to confirm these predictions with appropriate patient and phantom CT data in the future. Please note that the above generalizability discussion is regarding the acquisition parameters assuming that the body part to be scanned is the same. When a network is trained on CT images of the abdomen, it may not maintain the denoising performance in head or extremity scans and vice versa, since the noise property could differ significantly due to substantial changes in anatomical structure and size in a different body part.

A limitation of this work is that it investigated generalizability of a single denoising network, REDCNN. There are other popular networks applied to low-dose CT image denoising, such as ResNet, UNet and GAN. Different networks may have different ways of extracting relevant features in the training data, resulting in images of different resolution and noise properties [25]. However, DL methods share a common property: data-driven-based learning machnism. Therefore, training data is always an essential element affecting the performance of DL methods. We anticipate that the generalizability performances observed on REDCNN likely apply to other types of DL networks if they are similarly trained to perform a slice-wise low-dose CT image denoising function. The experiments conducted in this work will be performed using other typical types of DL networks to confirm this anticipation.

In summary, generalizability performance is an important characteristic of DL methods. Loss of generalizability of a DL network can be rooted in a shift of the testing data distribution from the training data. There are many different CT scan acquisition settings. Without any knowledge about the generalization behavior, we may have to test a CT image denosing network tediously on data from a large variety of scan settings to understand its use range. Our results imply that the noise correlation property described by NPS may be used as one way to predict the generalizability zone of a DL-based CT image denoising network. CT images with acquisition parameters that significantly change the NPS relative to the training data would possibly fall out of the generalizability zone, such as images reconstructed with a different convolutional kernel. CT images with acquisition parameters that have similar NPS shape to that of the training data would be still within the generalizability zone, such as the slice thickness parameter. This finding can be helpful to the development as well as regulartory evaluation of DL-based CT image denoising methods. For developers, the training data cohort may be more effectively designed. One may emphasize adding training data that has different NPS properties to improve the generalizability of a CT image denoising network or training the network separately on those categories of data. For regulartory evaluation, the categories of testing data may be appropriatedly reduced to support the assessment of the generalizability of a DL-based CT image denoising software within its intended use,

according to the FDA least-burdensome principle (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/least-burdensome-provisions-concept-and-principles). Validated intended uses and product labelings will allow clinicians to have better information on what kinds of images are suitable to be processed by a DL denoising algorithm available at their site.

5. CONCLUSIONS

This paper reported our work in testing the performance (MSE, MTF, NPS and LCD) generalizability of a DL-based CT denoising method (REDCNN) on three CT acquisition parameters (reconstruction kernel, slice thickness and dose). Our results showed that the DL performance did not generalize well between the sharp and smooth reconstruction kernels, was not sensitive to the slice thickness parameter, and was better when trained with mixed-dose data. The observed DL performance behaviors provide evidence to support the hypothesis that the noise property of training data, specifically the NPS, may be a data characteristic to predict the generalizability zone of a DL-based CT image denoising network. Future work is needed to investigate the impacts of other acquisition parameters on the performance generalizability to consolidate this hypothesis. Tasks that challenge possible differences in the higher spatial-frequency content of the denoised images should also be explored to allow a more complete performance evaluation.

REFERENCES

[1] M. Söderberg, and M. Gunnarsson, "Automatic exposure control in computed tomography--an evaluation of systems from different manufacturers," *Acta Radiol,* vol. 51, no. 6, pp. 625-34, Jul, 2010.

[2] S. M. Huck, G. S. K. Fung, K. Parodi, and K. Stierstorfer, "The z-sbDBA, a new concept for a dynamic sheet-based fluence field modulator in x-ray CT," *Medical Physics,* vol. 47, no. 10, pp. 4827-4837, 2020.

[3] I. A. Elbakri, and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," *IEEE Transactions on Medical Imaging,* vol. 21, no. 2, pp. 89-99, 2002.

[4] L. P. Qi, Y. Li, L. Tang, Y. L. Li, X. T. Li, Y. Cui, Y. S. Sun, and X. P. Zhang, "Evaluation of dose reduction and image quality in chest CT using adaptive statistical iterative reconstruction with the same group of patients," *The British journal of radiology,* vol. 85, no. 1018, pp. e906-e911, 2012.

[5] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network," *IEEE Transactions on Medical Imaging,* vol. 36, no. 12, pp. 2524-2535, 2017.

[6] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose CT via convolutional neural network," *Biomedical Optics Express,* vol. 8, no. 2, pp. 679-694, 2017/02/01, 2017.

[7] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss," *IEEE Transactions on Medical Imaging,* vol. 37, no. 6, pp. 1348-1357, 2018.

[8] B. Kim, M. Han, H. Shim, and J. Baek, "A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images," *Medical Physics,* vol. 46, no. 9, pp. 3906-3923, 2019.

[9] R. D. MacDougall, Y. Zhang, M. J. Callahan, J. Perez-Rossello, M. A. Breen, P. R. Johnston, and H. Yu, "Improving Low-Dose Pediatric Abdominal CT by Using Convolutional Neural Networks," *Radiology: Artificial Intelligence,* vol. 1, no. 6, pp. e180087, 2019.

[10] W. Yang, H. Zhang, J. Yang, J. Wu, X. Yin, Y. Chen, H. Shu, L. Luo, G. Coatrieux, Z. Gui, and Q. Feng, "Improving Low-Dose CT Image Using Residual Convolutional Network," *IEEE Access,* vol. 5, pp. 24698-24705, 2017.

[11] J. Solomon, P. Lyu, D. Marin, and E. Samei, "Noise and spatial resolution properties of a commercially available deep learning-based CT reconstruction algorithm," *Medical Physics,* vol. 47, no. 9, pp. 3961-3971, 2020.

[12] M. Lenfant, O. Chevallier, P.-O. Comby, G. Secco, K. Haioun, F. Ricolfi, B. Lemogne, and R. Loffroy, "Deep Learning Versus Iterative Reconstruction for CT Pulmonary Angiography in the Emergency Setting: Improved Image Quality and Reduced Radiation Dose," *Diagnostics (Basel, Switzerland),* vol. 10, no. 8, pp. 558, 2020.

[13] H. Kawashima, K. Ichikawa, T. Takata, W. Mitsui, H. Ueta, N. Yoneda, and S. Kobayashi, "Performance of clinically available deep learning image reconstruction in computed tomography: a phantom study," *J Med Imaging (Bellingham),* vol. 7, no. 6, pp. 063503, Nov, 2020.

[14] C. H. McCollough, A. C. Bartley, R. E. Carter, B. Chen, T. A. Drees, P. Edwards, D. R. Holmes III, A. E. Huang, F. Khan, S. Leng, K. L. McMillan, G. J. Michalak, K. M. Nunez, L. Yu, and J. G. Fletcher, “Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge,” *Medical Physics,* vol. 44, no. 10, pp. e339-e352, 2017.

[15] R. Zeng, C. Y. Lin, Q. Li, L. Jinag, J. A. Fessler, and K. Myers, “Generalizability test of a deep learning-based CT image denoising method,” *The 6th International Conference on Image Formation in X-Ray Computed Tomography*, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." pp. 770-778.

[17] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising,” *IEEE Transactions on Image Processing,* vol. 26, no. 7, pp. 3142-3155, 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* pp. 234-241.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672–2680, 2014.

[20] R. Zeng, S. Divel, Q. Li, and K. Myers, "Performance Evaluation of Deep Learning Methods Applied to CT Image Reconstruction." pp. E162-E162.

[21] S. Richard, D. B. Husarik, G. Yadava, S. N. Murphy, and E. Samei, “Towards task-based assessment of CT performance: System and object MTF across different reconstruction algorithms,” *Medical Physics,* vol. 39, no. 7Part1, pp. 4115-4122, 2012.

[22] J. Y. Vaishnav, W. C. Jung, L. M. Popescu, R. Zeng, and K. J. Myers, “Objective assessment of image quality and dose reduction in CT iterative reconstruction,” *Medical Physics,* vol. 41, no. 7, pp. 071904, 2014.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*: The MIT Press, 2016.

[24] C. Shorten, and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data,* vol. 6, no. 1, pp. 60, 2019/07/06, 2019.

[25] P. Kc, R. Zeng, M. Farhangi, and K. Myers, "Deep neural networks-based denoising models for CT imaging and their effcacy."