Received: 15 July 2021 | Revised: 22 November 2021 | Accepted: 8 December 2021

DOI: 10.1002/mp.15430

RESEARCH ARTICLE

MEDICAL PHYSICS

Performance of a deep learning-based CT image denoising method: Generalizability over dose, reconstruction kernel, and slice thickness

Rongping Zeng[1] | **Claire Yilin Lin**[2] | **Qin Li**[3] | **Lu Jiang**[1] | **Marlene Skopec**[4] | **Jeffrey A. Fessler**[5] | **Kyle J. Myers**[1]

[1] Center for Devices and Radiological Health, US Food and Drug Administration (FDA), Silver Spring, Maryland, USA

[2] KLA Corporation, Milpitas, California, USA

[3] AstraZeneca, Waltham, Massachusetts, USA

[4] National Institutes of Health, Bethesda, Maryland, USA

[5] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Rongping Zeng, Center for Devices and Radiological Health, US Food and Drug Administration (FDA), Silver Spring, MD 20993, USA.
Email: rongping.zeng@fda.hhs.gov

Abstract

Purpose: Deep learning (DL) is rapidly finding applications in low-dose CT image denoising. While having the potential to improve the image quality (IQ) 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. To study how the network generalizes on the three imaging parameters, we grouped the CT volumes in the Low-Dose Grand Challenge (LDGC) data into three pairs of training datasets according to their imaging parameters, changing only one parameter in each pair. We trained REDCNN with them to obtain six denoising models. We test each denoising model on datasets of matching and mismatching parameters with respect to its training sets regarding dose, reconstruction kernel, and slice thickness, respectively, to evaluate the denoising performance changes. Denoising performances are evaluated on patient scans, simulated phantom scans, and physical phantom scans using IQ metrics including mean-squared error (MSE), contrast-dependent modulation transfer function (MTF), pixel-level noise power spectrum (pNPS), and low-contrast lesion detectability (LCD).

Results: REDCNN had larger MSE when the testing data were 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 higher-dose images than that trained with mixed-dose data (17%–80%). 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 pNPS 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 arbitrary dose levels when trained with mixed-dose data, and not highly sensitive to slice thickness. It is known that reconstruction kernel affects the in-plane pNPS shape of a CT image, whereas slice thickness and dose level do not, so it is possible that the generalizability performance of this CT image denoising network highly correlates to the pNPS similarity between the testing and training data.

KEYWORDS
CT image denoising, deep learning, generalizability performance, image quality assessment

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 (IQ) to avoid misdiagnosis. Various approaches have been developed toward low-dose CT through improved hardware design such as automatic exposure control (AEC), 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 IQ over filtered back projection method (FBP), similar to the 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 predictable 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 is not straightforward. The term "generalizability" refers to the accuracy with which performance results can be transferred to situations or data other than those originally studied.[14] The generalizability zone then refers to the data range for which a DL method preserves its reference performance, which is usually achieved when the testing and training data are acquired under the same condition. Preserving the performance means that the performance tested on a new set of data is comparable or statistically equivalent to the reference performance. The performance can be multifaceted for an image reconstruction and denoising method depending on the specifications. We considered multiple IQ metrics in this work as described in the next paragraph. In CT, image resolution and noise properties are affected by CT imaging parameters including both the raw data acquisition parameters (kVp, mA, collimation width, pitch, etc.) and the reconstruction parameters (reconstruction kernel, slice thickness, reconstruction field of view, etc.). Therefore, it is reasonable to investigate the generalizability performance of a DL network on data acquired with different parameters. Changes in the network's performance when tested on differently acquired datasets could indicate a potential data distribution shift caused by the associated imaging parameters. Thus, an analysis of the data properties associated with the imaging 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[15] to train that network.[16] We examined the denoising performance changes between the conditions of training/testing using data with matching and mismatching imaging parameters under three scenarios. In each scenario, only one imaging parameter changed between the training and testing data. The three imaging parameters were reconstruction kernel, slice thickness, and dose level. The 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 pixel-level noise power spectrum (pNPS), standard

CT IQ metrics that characterize the image resolution and noise properties (pNPS differs from the standard NPS only in terms of the dimensional unit, as explained in Section 2.3.2); 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 IQ. 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.

A similar study was conducted by Huber et al. that evaluated the performance of one narrowly trained denoising network on processing images reconstructed differently from the training data in terms of field of view (FOV), reconstruction kernel, and slice thickness.[17] It was observed in the Huber et al. study that the denoising performance was degraded with variations in FOV and kernel, but not affected by thickness. We also evaluated the performance behavior of a DL denoising network in matching and mismatching test data. However, a different denoising network was examined and the training and testing conditions were not designed the same. In this sense, our study and the Huber et al. study are complementary to each other. In addition, the IQ evaluation methods in our work, including MTF, pNPS, and LCD, are more comprehensive than those used in the Huber et al. study. Furthermore, we analyzed the underlying training and testing data properties to help answer the question of why some parameters may cause a data distribution shift to affect the generalizability and some may not. As is known, imaging 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. The dose level determined the noise magnitude. A degradation in the DL network's denoising efficiency due to a mismatch in an imaging parameter may be associated with a shift of the underlying data properties that are caused by that parameter. Based on the observations regarding whether a change in each of the three parameters causes a substantial degradation in the DL's denoising performance or not, we may learn and identify which underlying data properties are most important in predicting the denoising network's generalizability. Note that CT image noise is also object-dependent. A DL network trained with body scans may not generalize well to scans of other anatomy such as head scans or extremity scans even if the CT imaging parameters are kept the same. This study focused on the impact of CT imaging parameters since our training data are exclusively from the body scans covering the segment from thorax to abdomen.

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 $x \in R^{m\times n}$ denote a low-dose CT reconstructed image; the DL-based denoising problem is to optimize the network $C(x): R^{m\times n} \rightarrow R^{m\times n}$ that maps x to its corresponding high-dose image $y \in R^{m\times n}$ by minimizing a loss function between x and 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,[5,10,18,19] UNet,[8,20] and generative adversarial networks.[7,21] For this paper, we selected the RED-CNN 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 Figure 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 the Chen et al. paper.[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.[5,22]

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.

In our implementation, the denoising network was trained using many pairs of two-dimensional (2D) small image patches extracted from low-dose and corresponding full-dose patient CT slices, as described below in Section 2.2. Therefore, it was a 2D denoising network. After the network is trained, it can be applied directly to

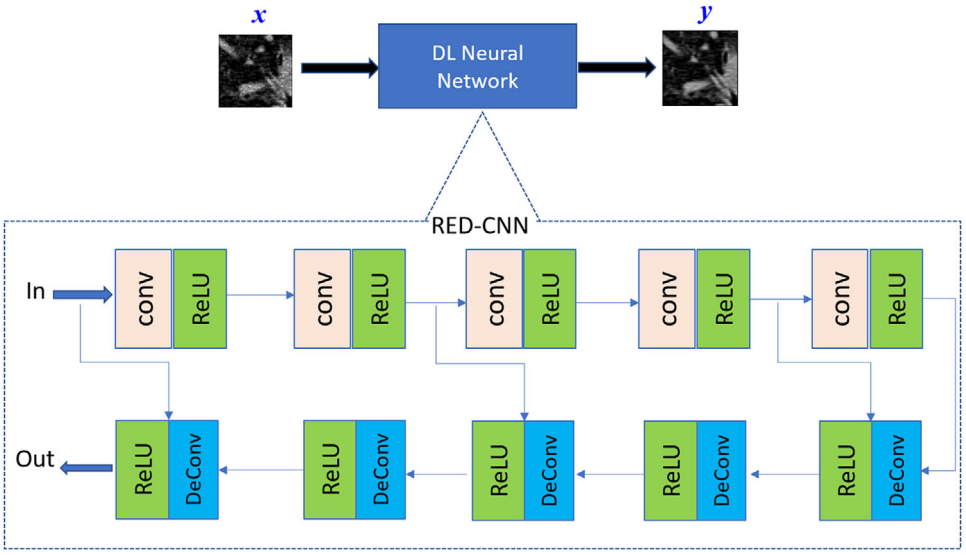


FIGURE 1 Illustration of the residual encoder–decoder convolutional neural network (REDCNN) denoising network. ***x*** and ***y*** represent the noisy image input and the cleaner image output. Each "conv" layer contains 96 filters of size 5 × 5. Each "DeConv" layer also contains 96 filters of size 5 × 5 except that the last DeConv has only one filter of size 5 × 5

a whole image slice since the "conv" and "DeConv" functions in the neural network are essentially convolution operations.

2.2 | Training data categorization

The denoising network was trained using the patient scans in the LDGC dataset.[15] 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. Three hundred fifty slices of size 512 × 512 were randomly selected from the seven patients and each slice was divided into 55 × 55 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 imaging parameters as shown below. In each pair, only one imaging 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–3 mm–mix%" represents the REDCNN trained with images of sharp kernel, 3 mm thickness, and mixed dose levels; "DLsmooth–1 mm–25%" represents the REDCNN trained with images of smooth kernel, 1 mm 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:

$$\mathbf{x_d} = \mathbf{x_f} + \alpha\left(\mathbf{x_q} - \mathbf{x_f}\right),\ \alpha \geq 0, \tag{1}$$

where $\mathbf{x_d}$, $\mathbf{x_f}$, and $\mathbf{x_q}$ represent the synthesized noisy image at a dose level d, the original full-dose, and the quarter-dose images, respectively. The scalar 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. Assuming the full-dose and quarter-dose image noise variances are σ^2 and $4\sigma^2$ at an image pixel, for an arbitrary nonnegative blending factor α, the noise variance of $x_d$at the same pixel will be $((1-\alpha)^2 + 4\alpha^2)\sigma^2$. The noise level corresponds to $1/((1-\alpha)^2 + 4\alpha^2)$ of the full-dose scan, based on the relationship that noise variance is inversely proportional to the exposure level in CT images reconstructed with FBP when all the other scan parameters are the same. We varied the blending factor randomly in the interval of [0.5, 1.2] for the mixed-dose training data case, resulting in images corresponding to 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, pNPS, and LCD. MSE reflects how well the network performs in minimizing the loss function that it is designed to do. We did not evaluate 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 IQ. 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.

2.4 | Mean-squared error 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 3 mm and 1 mm slice thickness CT volumes, respectively. For each slice, the full-dose image was used as a reference to calculate the MSE (=||Noisy image – Ref image||^2/The total number of pixels) before and after the DL denoising. Then, the MSE reduction rate (=[(MSE before denoising – MSE after denoising)/MSE before denoising] × 100%) was calculated to quantify the denoising performance. The MSE was evaluated on an entire image slice. 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.

2.5 | Contrast-dependent modular transfer function and pixel-level noise power spectrum test

We simulated 2D phantom CT scans for the MTF and NPS tests. We also collected physical CT scans of the CATPHAN600 (The Phantom Laboratory, Salem, NY) to validate the simulation-based results, which are described later in Section 2.4. Both the MTF and NPS were evaluated within the plane. We did not evaluate the z-directional resolution and noise property because REDCNN was implemented as a 2D denoising network in this study. The network model was not trained using the z-direction data and hence was not designed to alter the z-directional property in an image slice. The simulated contrast phantom had a similar layout as the CATPHAN600 contrast module CTP404 to allow the measurement of contrast-dependent MTF. In this work, the contrast-dependent MTF was evaluated using the methods described in the Richard el al. paper[23] at these five contrast levels: 990, 340, 200, 120, and 35 HU. The MTF50% value was recorded for each MTF curve and plotted as a function of the HU contrast to characterize the contrast-dependent image resolution.

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. However, we also simulated five noisy contrast phantom scans for the MTF test to validate that using a noiseless scan to assess the resolution property of a nonlinear DL noise reduction is appropriate. For the NPS measurement, 50 noisy water phantom CT scans were simulated. A region of interest (ROI) of size 64 × 64 pixels at the image center was extracted from each realization. Local noise power spectrum (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. A one-dimensional (1D) NPS curve was also estimated by radially binning the corresponding 2D NPS image. Because the DL network was trained to operate on pixelized images without being informed about the length unit of the pixel size, the estimated NPS was considered to be a function of the discrete frequency unit “cyc/pix” (cycle/pixel) in this study, rather than “lp/cm” (line pair/cm). To differentiate from the standard NPS that usually has a dimensional unit of “lp/cm”, we refer to the “cyc/pix” unit-based NPS as pixel-level NPS, shortened as pNPS.

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

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

to achieve different noise levels. To simulate the reconstruction kernels in the LDGC data, two Hann filters of different cutoff frequencies (named Hann1 and 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 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.

2.5.1 | Low-contrast detectability test

The low-contrast detectability was estimated using a model observer and simulated MITA-LCD phantom (The Phantom Laboratory, Salem, NY) CT images. Specifically, we simulated 200 CT scans of the signal module and 100 scans of the background module of the MITA-LCD phantom CCT189 (Figure 2) at five exposure levels. The signal module contained four low-contrast disks with varying size/HU combinations (3 mm/14HU, 5 mm/7HU, 7 mm/5HU, 10 mm/3HU) to mimic subtle lesions. The five exposure levels we simulated were: 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.[24] 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).

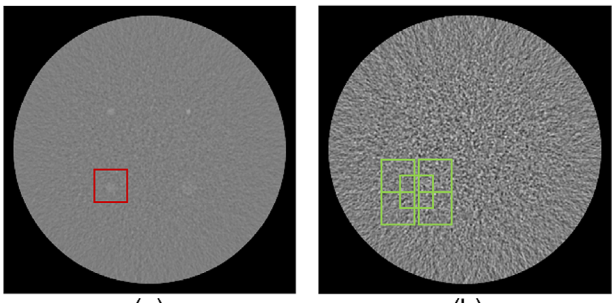


FIGURE 2 Sample CT images of the simulated MITA-LCD phantom signal module (left) and background module (right) for the LCD test. Red and green boxes illustrate the locations for cropping signal-present region of interests (ROIs) and the corresponding signal-absent ROIs. Note that the CT image of the signal module shown here is an average of 20 realizations from the highest dose level reconstructed with filtered back projection (FBP) of smooth kernel to make the low-contrast signals visible. The display window is [−50 50] for both images

2.6 | 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 FOV. 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 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

TABLE 2 Comparison of the imaging 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)	0.6/38.4	0.6–0.8	378 (mean)	3	B30f
				WEDGE_3 (2)				1	D45f
Phantom scans	OFF	120	32.1 (High)	FLAT	0.6/12	0.8	380	3	B30f
			20.0 (Full)					1	D45f
			5.0 (Quarter)						

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 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 380 mm, 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 3 mm slice thickness, sharp and smooth kernel were generated, resulting in 44 CT volumes.

3 | RESULTS

3.1 | Mean-squared errors test

Figure 3 shows the plots that compare the MSE reduction rates of the three pairs of DL networks.

For the dose effect (Figure 3a), when tested on the quarter-dose images, the DL networks trained solely with quarter-dose data (DLsmooth_3 mm_25%) and trained with mixed-dose data (DLsmoooth_3 mm_mix%) had almost equivalent MSE reduction rate. When tested on the 80% dose images, the DL network trained with mixed dose reduced MSE noticeably more. We also tested the network models on another two dose levels: 50%, a moderate low-dose level and 18%, close to the lowest dose level in the mixed-dose training data. Difference of MSE reduction rate between the two models over the four tested dose levels is plotted in Figure 4. It can be seen that DLsmoooth–3 mm–mix% had similar MSE performance at aggressively low dose levels (18% and 25%) compared to DLsmoooth–3 mm–25% but reduced MSE comparatively more as the dose level increased toward normal dose. This trend indicates that the DL denoising network trained with mixed-dose data generalized better on data of different dose levels.

For the reconstruction kernel effect, Figure 3b shows that when the training and testing data had a different reconstruction kernel, the DL network performed substantially worse than the cases with matching reconstruction kernel in the training and testing data. This indicates that the DL denoising network did not generalize well on data with a different reconstruction kernel.

For the thickness effect (Figure 3c), in both the 3 mm and 1 mm 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 3 mm 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.

Figure 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 Figure 5), the DLsharp–3 mm–mix% processed image obviously appears to be much noiser than the image processed by DLsmooth–3 mm–mix%. Meanwhile, in the test case of FBP sharp (bottom two rows in Figure 5), the image texture of the DLsmooth–3 mm–mix% processed FBP sharp image appears quite different from the others. It is also noticeable that the anatomical structures in the DLsmooth–3 mm–mix% processed image slice are oversmoothed and some small features are lost.

3.2 | Contrast-dependent modular transfer function test

Figure 6a and b shows the contrast-dependent image resolution curves for the DL networks evaluated using the simulated noiseless FBP-smooth and FBP-sharp contrast phantom images, respectively. The curves clearly show that the image resolution decreases with contrast. This nonlinear smoothing behavior is similar

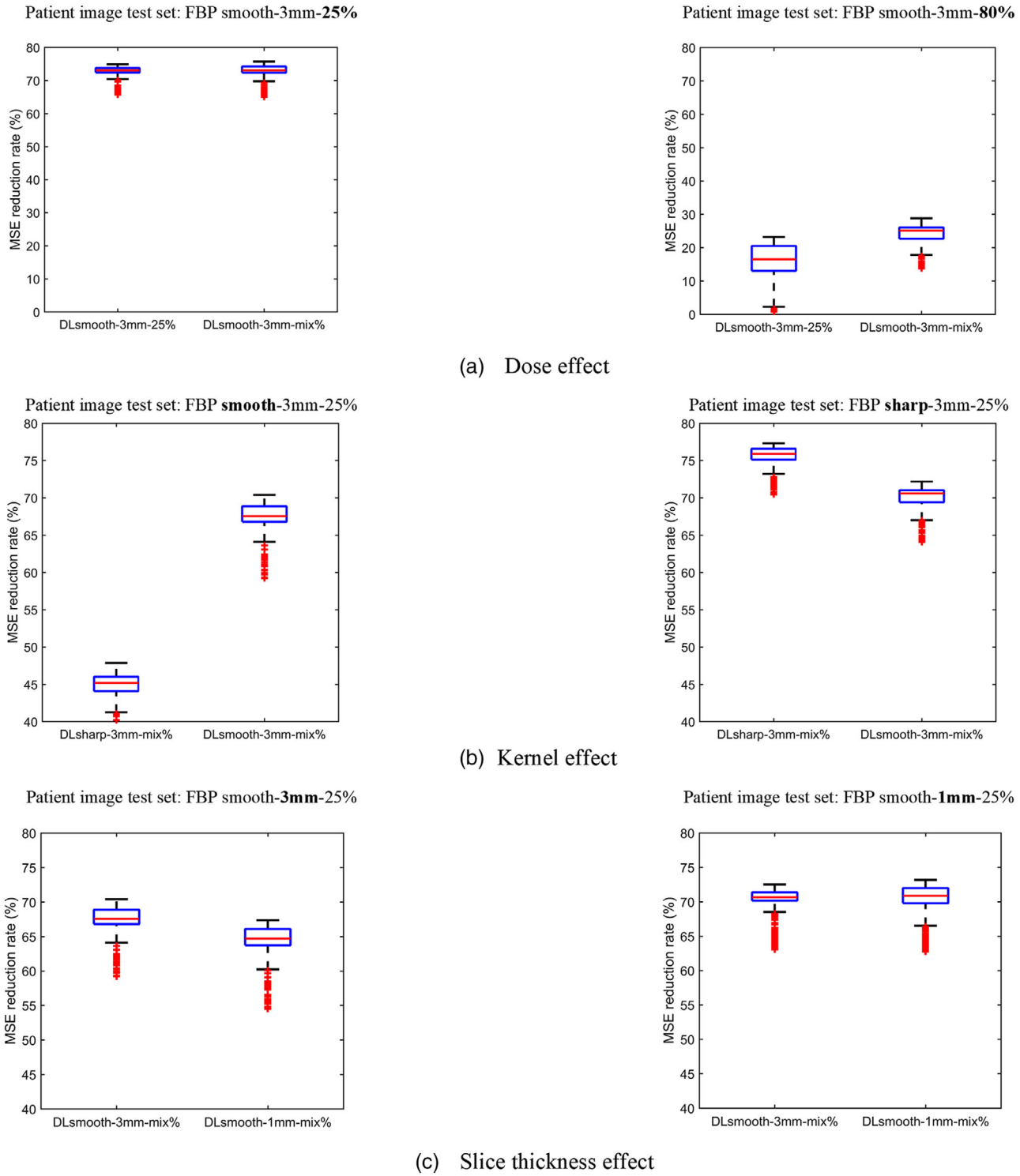


FIGURE 3 Plots of the mean-squared error (MSE) reduction rate of the deep learning (DL) networks tested on patient images of matching and mismatching CT imaging parameters. (a)The first row compares the dose level effect on denoising models trained with single-dose (DLsmooth–3 mm–25%) and mixed-dose data (DLsmooth–3 mm–mix%). (b) The second row compares the reconstruction kernel effect on denoising models trained with sharp kernel data (DLsharp–3 mm–mix%) and smooth kernel data (DLsmooth–3 mm–mix%). (c) The third row compares the thickness effect on denoising models trained with 3 mm thickness data (DLsmooth–3 mm–mix%) and 1 mm thickness data (DLsmooth–1 mm–mix%). The box plots were generated using the Boxplot() function in MATLAB, in which the central red line indicates the median and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the "+" marker symbol

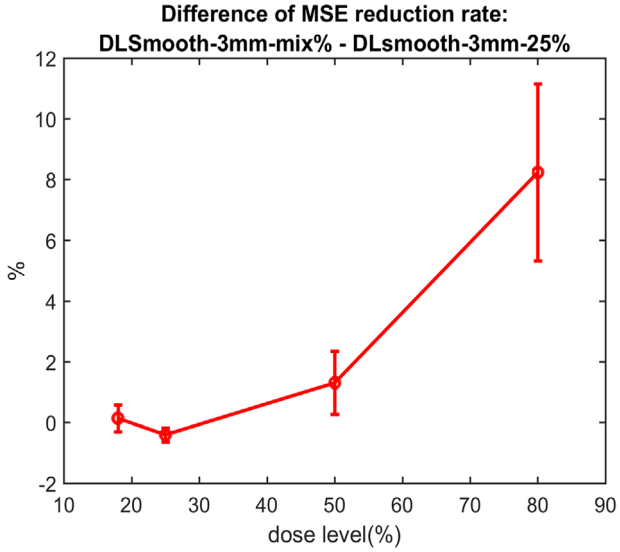


FIGURE 4 Difference of mean-squared error (MSE) reduction rate between the deep learning (DL) model trained with mixed dose data (DLsmooth–3 mm–mix%) and the DL model trained with quarter-dose data (DLsmooth–3 mm–25%). Test datasets were patient CT images reconstructed with smooth kernel and 3 mm slice thickness at four dose levels: 18%, 25%, 50%, and 80%

to that of traditional IR and denoising methods. It is also observed from Figure 6a,b that relative MTF performance among the differently trained network models are the same in the tests of using FBP-smooth and FBP-sharp contrast phantom images: 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 quarter-dose data, except at the contrast level of 35HU, where the resolution dropped greatly for the quater-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 better preserved.

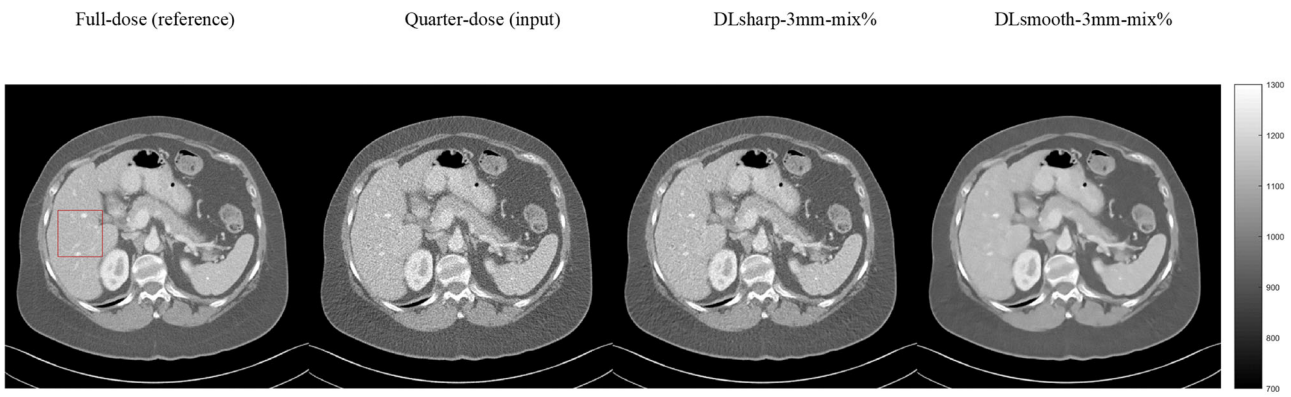
To validate that the MTF measurements evaluated using noiseless test image are consistent with the results obtained on noisy test images, we also simulated five noisy contrast phantom images with an air photon count of 2.4×10^5 per pixel and further tested the MTF performance of one DL network model (DLsmooth–3 mm–mix%). Figure 6c compares the MTF50% curves of DLsmooth–3 mm–mix% in noiseless and noisy test images. It can be seen that the mean of MTF50% measurements using noisy test images were consistent with those on a noiseless test image, but with additional uncertainty from image noise.

We also checked whether the DL network would introduce bias to the HU values of the contrast objects when tested on noiseless image. Using the DLsmooth–3 mm–mix% model as an example, Table 3 lists the mean HU values (over the central 9 pixels) of the disks in the original FBP contrast-phantom images and the DL processed images, for both noiseless and noisy cases. The HU values in the DL processed noiseless and noisy FBP images were all close to the simulated true values. The HU results show that no bias was introduced by the DL model when tested on noiseless FBP images although it was trained using noisy CT images.

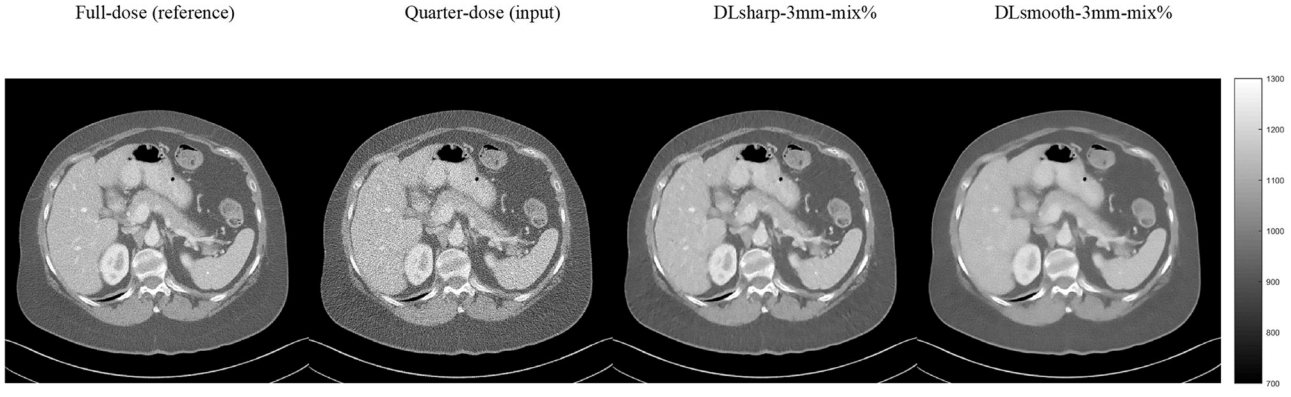
3.3 | Pixel-level noise power spectrum test

We simulated 50 noisy scans of a cylindrical water phantom for the pNPS estimation, with an air photon count of 2.4×10^5 per pixel. Each noisy scan was reconstructed by FBP with both sharp and smooth kernels. Then, the noisy images were processed by DLsharp–3 mm–mix% and DLsmooth–3 mm–mix% 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 previous MSE and MTF test results showed that the DL network trained with 3 mm slice thickness and mixed-dose data had better performances.

Figure 7 presents the local pNPS images and Figure 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 IR and denoising methods. In general, DL denoised images had noise components concentrated more in the lower frequency bands compared to the original FBP images. The local pNPS of the DL images also appear to be less radially isotropic, reflecting higher nonstationarity along the angular direction of the DL noise reduction method. In addition, one may notice a contrasting appearance in the pNPS of DLsmooth–3 mm–mix% processed FBP-sharp image (the rightmost in Figure 7b): much higher magnitude at the four corners (high-frequency regions). The 1D radial profile clearly shows that the corresponding pNPS curve has a rising tail (as indicated by the arrow in Figure 8b) after about 0.5 cyc/pix. Moreover, the tail's shape and height closely match those of the pNPS curve of the original FBP-sharp images, indicating that the high-frequency noise was not removed by the DL network trained with smooth-kernel data. An image patch of the DLsmooth–3 mm–mix% denoised FBP-sharp water phantom image, shown in Figure 8c, also demonstrates the remaining high-frequency noise, appearing as tiny checkerboard like artifacts. This phenomenon suggests that the DL model possibly did not learn to remove the high-frequency noise from the



(a) Denosing example using a patient image reconstructed with smooth kernel and 3 mm slice thickness



(b) Denosing example using a patient image reconstructed with sharp kernel and 3 mm slice thickness

FIGURE 5 Images to illustrate the effect of reconstruction kernel on the denoising models. From left to right are images of a full-dose filtered back projection (FBP) image as the reference, the quarter-dose FBP image as the input to the deep learning (DL) networks, DLsharp–3 mm–mix% and DLsmooth–3 mm–mix% denoised quarter-dose images. (a) Tested on a patient image slice reconstructed with smooth kernel and 3 mm slice thickness and (b) tested on a patient image slice reconstructed sharp kernel and 3 mm slice thickness. The red box in the full-dose FBP image in (a) indicates the region of interest (ROI) that is zoomed for display

TABLE 3 Comparison of the HU values of the five disks evaluated using simulated contrast phantom test images reconstructed with smooth kernel. In this table, the DL model that was used to process the contrast phantom images was DLsmooth–3 mm–mix%. Each HU value was calculated as an average over the central 9 pixels of the corresponding disk. For the case of noisy images, the two values before and in the parenthesis represent the mean and standard deviation, respectively, estimated from 5 noisy realizations. The HU results in the table show that no bias was introduced in the denoised images when tested on noiseless FBP images

Disk object	990 HU	340 HU	−200 HU	120 HU	−35 HU
Noiseless FBP-smooth image (reference)	989.9	339.7	−200.0	119.9	−35.0
DL processed noiseless FBP-smooth image	990.7	340.1	−199.9	120.0	−35.3
Noisy FBP-smooth images	989.1 (2.31)	341.1 (3.89)	−200.1 (3.37)	120.2 (2.48)	−32.8 (3.25)
DL processed noisy FBP-smooth images	989.9 (1.62)	341.0 (2.34)	−200.2 (2.38)	120.5 (1.46	−33.6 (2.02)

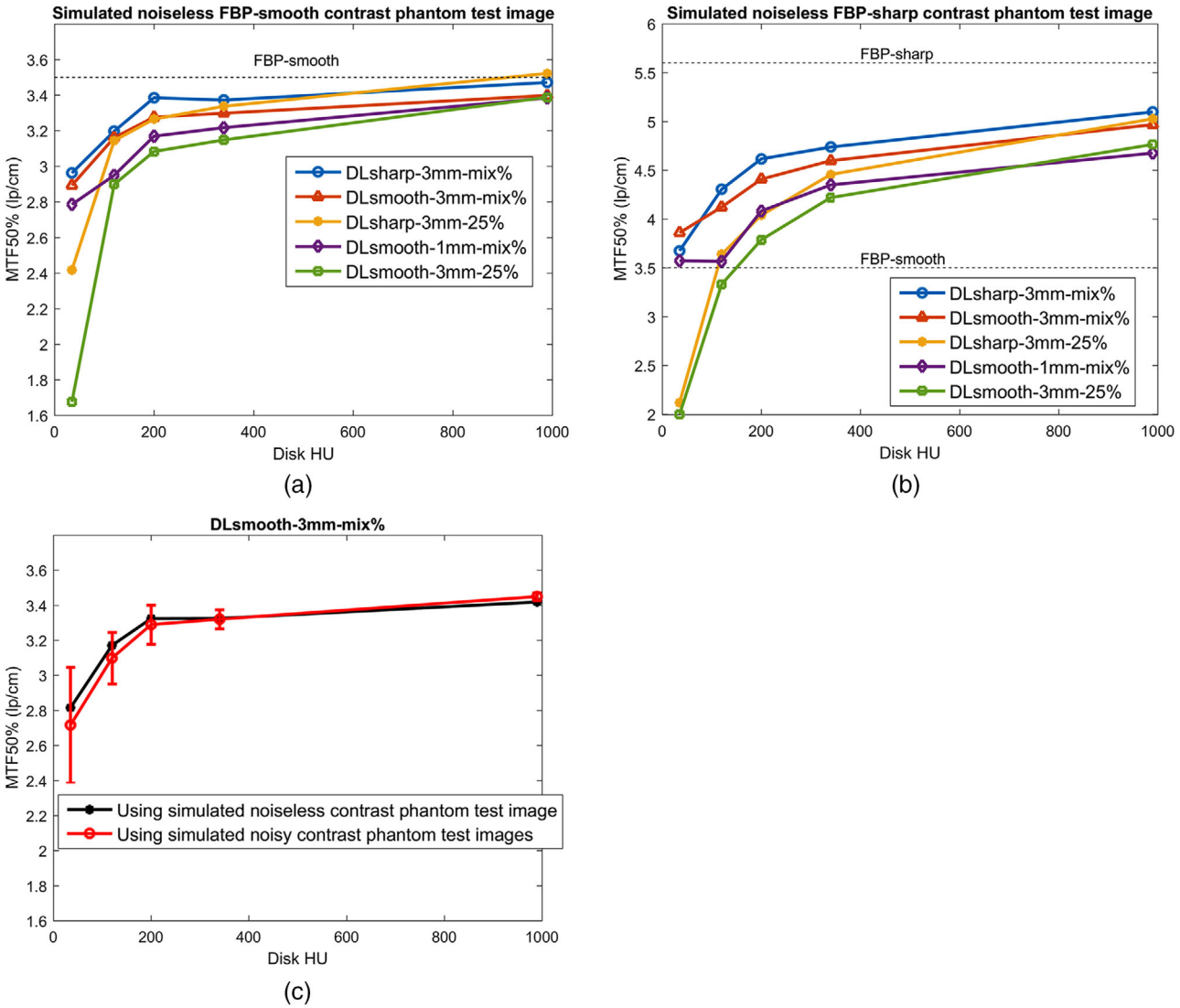


FIGURE 6 Contrast-dependent MTF50% curves of the DL networks evaluated using (a) the simulated noiseless contrast phantom test image reconstructed with smooth kernel and (b) the simulated noiseless contrast phantom test image reconstructed with sharp kernel. (c) Comparison of contrast-dependent MTF50% curves of the DLsmooth–3 mm–mix% model in noiseless and noisy contrast phantom test images reconstructed with smooth kernel

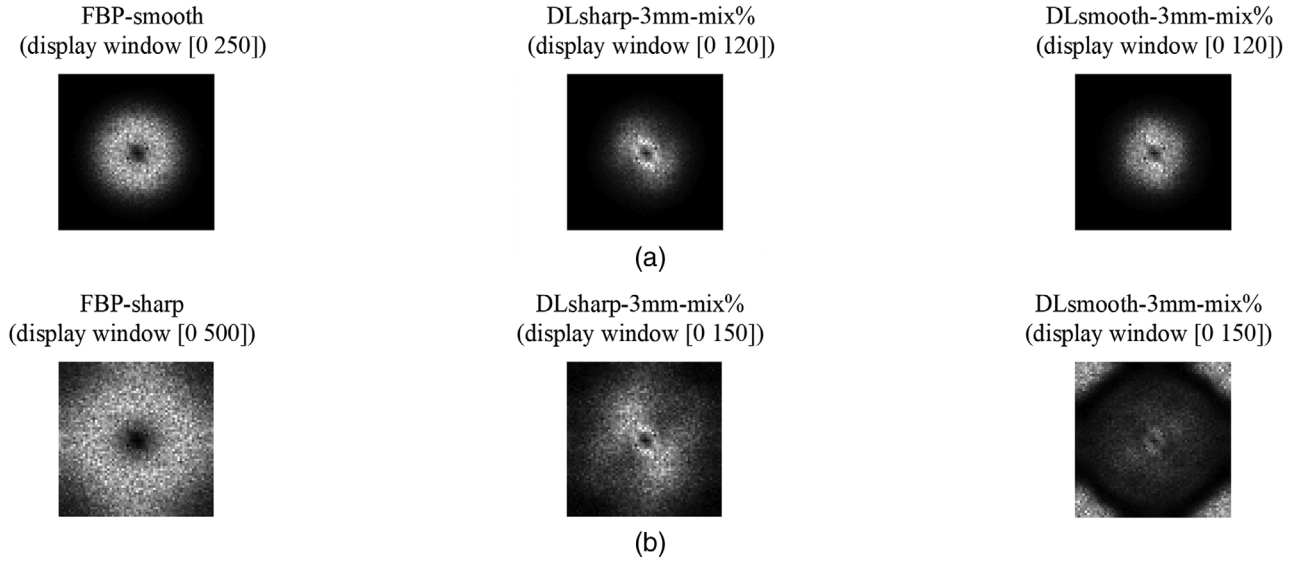


FIGURE 7 Two-dimensional (2D) local pixel-level noise power spectrum (pNPS) of the simulated water phantom images and the corresponding DLsharp–3 mm–mix% and DLsmooth–3 mm–mix% denoised images evaluated using: (a) simulated 2D water phantom images reconstructed with smooth kernel and (b) simulated 2D water phantom images reconstructed with sharp kernel

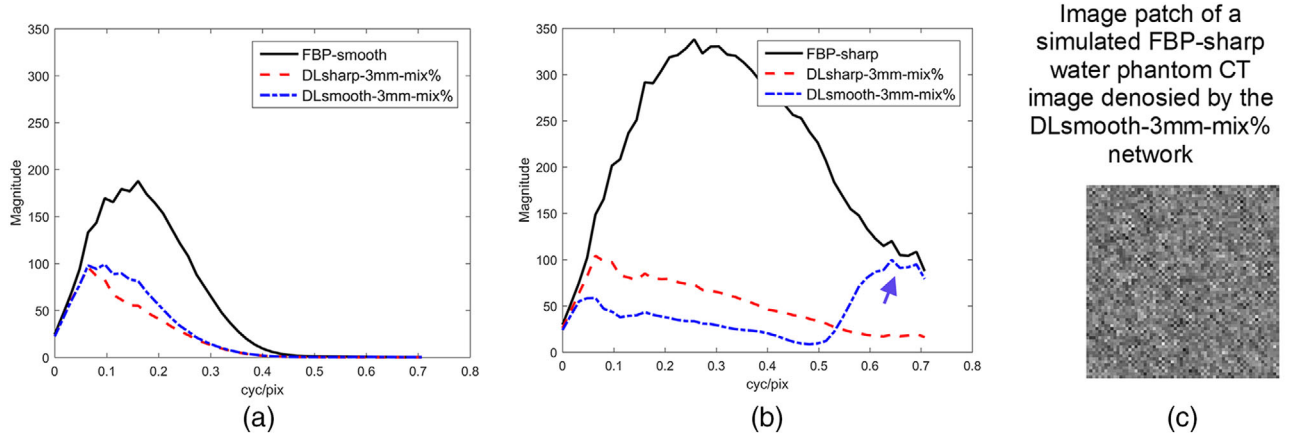


FIGURE 8 (a) The one-dimensional (1D) radial profiles of the pixel-level noise power spectrum (pNPS) images in Figure 7a. (b) The 1D radial profiles of the pNPS images in Figure 7b. The blue arrow in (b) points to the remaining high-frequency noise in the DLsmooth–3 mm–mix% processed filtered back projection (FBP)-sharp images. The sample image patch in (c) illustrates the remaining high-frequency noise, which appears as tiny checkerboard like artifacts

smooth kernel training data, since the training data did not contain noise in the high-frequency band.

3.4 | MTF and pNPS 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 3 mm-thickness and high-dose FBP images. Figure 9 displays the resolution curves. Due to image noise, the MTF function estimated from the disks of contrast below 100HU was 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 Figure 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 local pNPS images and extracted their 1D radial profiles of the DL networks processing 3 mm-thickness and full-dose FBP images,

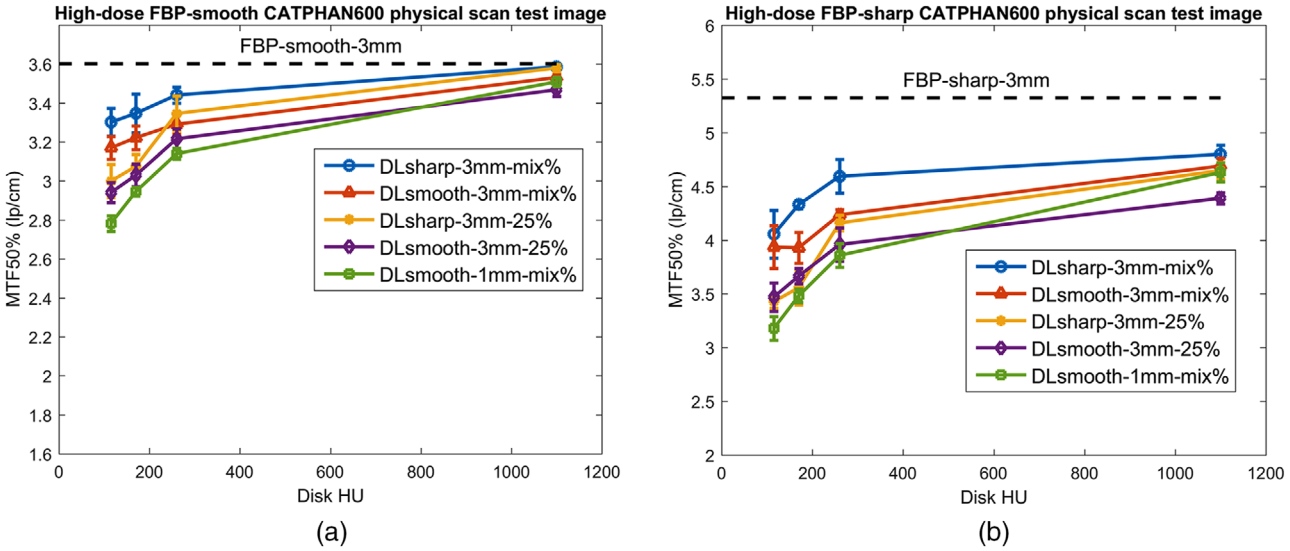


FIGURE 9 Contrast-dependent MTF50% curves of the deep learning (DL) networks evaluated using high-dose and 3 mm-slice-thickness CATPHAN600 physical CT test images reconstructed with (a) smooth kernel and (b) sharp kernel

shown in Figure 10. A rising tail in the 1D pNPS curve of the DLsmooth–3 mm–mix% processed FBP-sharp images was also observed, similar to that in Figure 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.

3.5 | Low contrast detectability test

Figure 11 plots AUC, a measure of low-contrast detectability, as a function of dose for detecting the 10 mm/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 (3 mm/14HU, 5 mm/7HU, 7 mm/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 (3 mm/14HU and 5 mm/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 10 mm/3HU insert since the curves separated the most in this case.

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 imaging 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, pNPS, 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, consistent with the observations in.[17] This is reasonable since the reconstruction kernel is the most dominant factor that determines the noise correlation structure in a FBP reconstructed image. The pNPS curves of the FBP-sharp and FBP-smooth images in Figures 8 and 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, consistent with the observations in.[17] Usually for FBP reconstructed volume in a helical CT scan, the slice thickness parameter is related to the

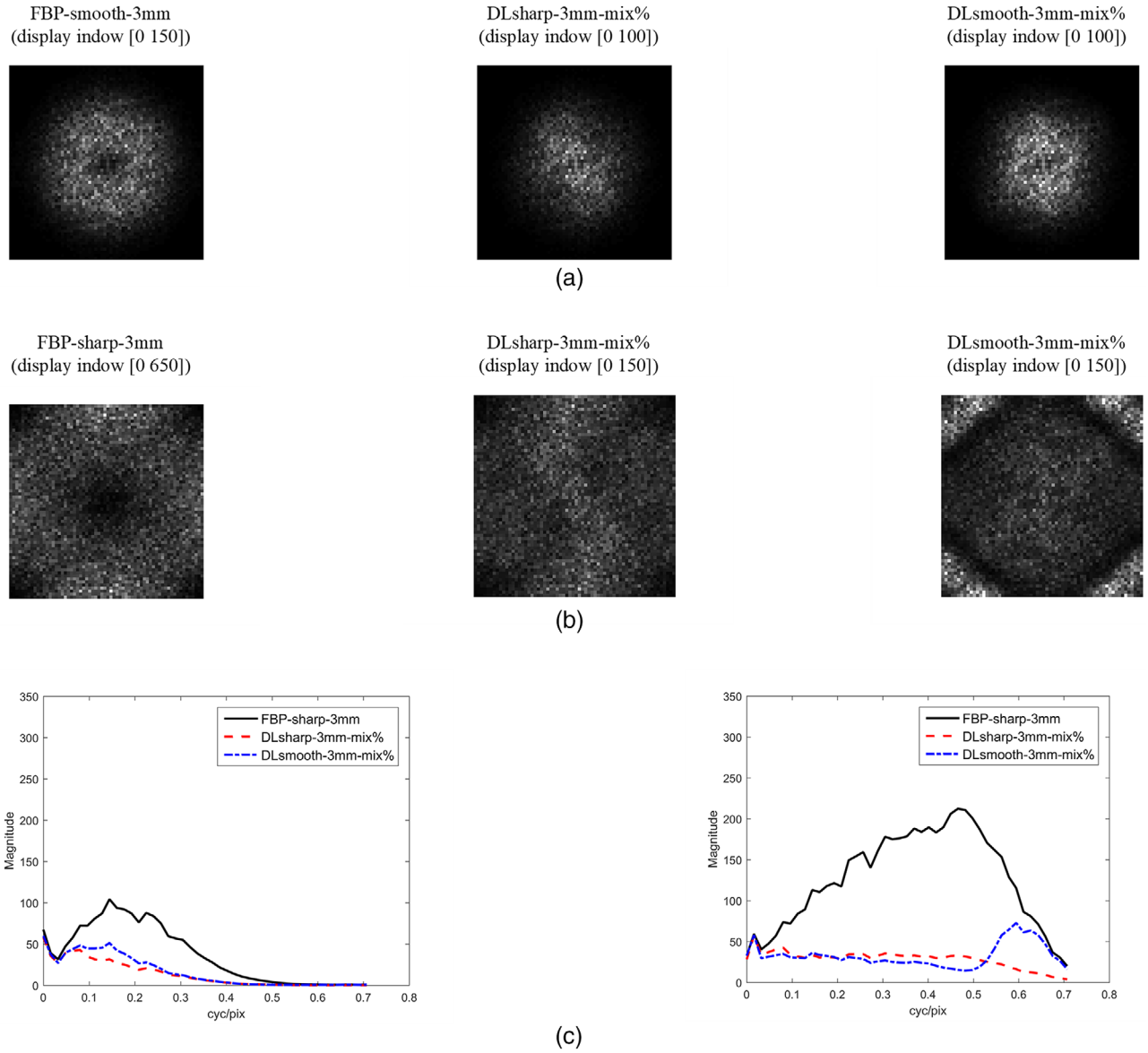


FIGURE 10 Two-dimensional (2D) local pixel-level noise power spectrum (pNPS) of the CATPHAN600 physical CT scans and the corresponding DLsharp–3 mm–mix% and DLsmooth–3 mm–mix% denoised images evaluated using (a) the CATPHAN600 images reconstructed with smooth kernel and 3 mm slice thickness, and (b) the CATPHAN600 images reconstructed with sharp kernel and 3 mm slice thickness. (c) Radial profiles of the 2D pNPS images in (a) on the left and the radial profiles of the 2D pNPS images in (b) on the right

interpolation width along the *z*-direction applied in the image reconstruction process.[25] When all the other imaging parameters are kept the same, a 3 mm slice thickness CT volume may be conceptually considered as being formed by a moving average (or weighted average) of adjacent slices of the 1 mm slice thickness CT volumes. Averaging along the *z*-direction does not alter the noise correlation structure within a slice, so the denoising networks trained with 3 mm and 1 mm thickness image slices were not much different. However, the noise magnitude in a 3 mm thickness slice is usually lower than that in the corresponding 1 mm slice. In this sense, the target images in the 3 mm thickness training data had slightly better IQ, which may explain why the DL–3 mm network performed slightly better than the DL–1 mm 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

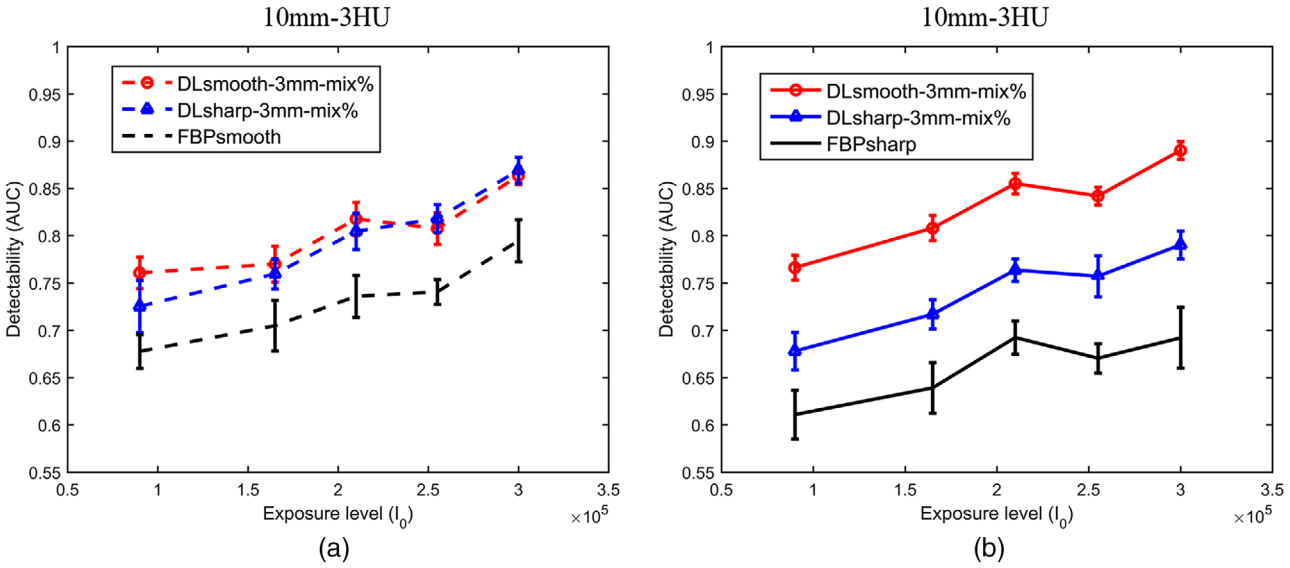


FIGURE 11 Detectability curves for the insert (10 mm–3HU) in the original noisy filtered back projection (FBP) images, and denoised FBP images with DLsharp–3 mm–mix% and DLsmooth–3 mm–mix% evaluated using (a) simulated MITA-LCD phantom images reconstructed with smooth kernel, and (b) simulated MITA-LCD phantom test images reconstructed with sharp kernel

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 the robustness of a DL network performance.[26,27]

Despite the finding based on the MSE and pNPS 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 7 mm and 10 mm disks for 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 Figure 7b, the rising tail of the NPS curve of DLsmooth starts at about 0.5 cyc/pix. Even for the smallest 3 mm disk that was about 4.5 pixel wide, its main spectrum lobe is within 0.22 cyc/pix; the signal power of most of the low-contrast disks included in the MITA-LCD phantom already diminishes at 0.5 cyc/pix. Based on the MTF and pNPS tests, the DLsmooth network model appeared to have comparable resolution and better noise reduction in the lower frequency band compared to the DLsharp network model, which may have contributed to the higher detectabilities of the DLsmooth network model. 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 imaging parameters in this work. Other parameters associated with a CT scan can also affect the FBP IQ, 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 an 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 evidence to support that the generalizability performance of REDCNN denoising algorithm is highly correlated with the in-plane pNPS property of the data determined by the CT imaging parameters: if a different imaging parameter value associated with the testing data does not alter the pNPS 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 pNPS shape, the DL network will likely have poorer denoising performance, such as between the sharp and smooth reconstruction kernels. Based on this finding, 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 may not be very 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, that is, 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 pNPS 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. Loss of resolution or denoising performance due to a change of FOV in the data was observed in the Huber et al. study, where a range of FOV from 100 mm to 400 mm was examined.[17] 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 imaging parameters assuming that the body part to be scanned is the same. When a network is trained on CT images of abdomen, it may not maintain the denoising performance in head or extremity scans and vice versa, since the object-dependent CT noise property could differ significantly due to substantial changes in anatomical structure and size of a 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.[28] However, DL methods share a common property: data-driven-based learning mechanism. Therefore, training data are 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 2D DL denoising networks if they are similarly trained to perform a slicewise CT image denoising function. The experiments conducted in this work will be performed using other typical types of DL networks to confirm this anticipation.

Another point that should be noted is the open question of the utility of MTF and NPS for the characterization of DL-derived images. These two Fourier-based IQ metrics are designed for linear, stationary systems (shift-invariant). FBP is a linear reconstruction method. An FBP-reconstructed CT image is approximately locally stationary,[29] hence MTF and NPS are suitable to describe a CT system with FBP reconstruction. Therefore, we used MTF and local NPS to analyze the underlying data property of the FBP image input to the denoising network. However, these metrics cannot fully characterize the imaging performance of a nonlinear noise reduction method.[24] A DL-based denoising process is obviously a nonlinear process. Modified versions of the MTF and NPS, such as contrast-dependent MTF[23] and noise-level dependent NPS,[30] have been proposed as possibly providing a better picture of the imaging performance of those nonlinear methods, but they may still not capture shift-variant properties of images. In this work, we evaluated the contrast-dependent MTF and local NPS for comparing the resolution and noise behavior over the imaged frequency range of various denoising network models measured under the same context. However, metrics like MTF and NPS developed for linear systems need to be complemented by other performance metrics such as task-based performance metrics to fully characterize the diagnostic image performance of DL methods.

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 imaging settings. Without any knowledge about the generalization behavior, we may have to test a CT image denoising network tediously on data from a large variety of scan settings to understand its use range. Our results imply that comparing the underlying pNPS associated with the imaging parameters used to acquire the testing and training data may be used as one way to predict the generalizability performance of a DL-based CT image denoising network. CT data acquired with imaging parameters that significantly change the pNPS relative to the training data would possibly not benefit from a DL noise reduction model, such as images reconstructed with a different reconstruction kernel. This finding can be helpful to the development as well as regulatory evaluation of DL-based CT image denoising methods. For developers, the training data cohort may be more effectively designed. One may emphasize on adding training data that has different pNPS properties to improve the generalizability of a CT image denoising network or training the network separately on those categories of data. For regulatory evaluation, the categories of testing data may be appropriately 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 labeling 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, pNPS, and LCD) generalizability of a 2D DL-based CT denoising method (REDCNN) on three CT imaging 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 highly sensitive to the slice thickness parameter, and was better when trained with mixed-dose data. The observed DL performance behaviors indicate that the generalizability performance of a DL-based CT image denoising network highly correlates to the pNPS similarity between the testing and training data. Future work is needed to investigate the impact of other imaging parameters on the performance generalizability to consolidate this finding. 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.

ACKNOWLEDGMENTS

Lin CY was affiliated with the University of Michigan-Ann Arbor and was supported in part by FDA Critical Path funding when she was involved in this work. Li Q was affiliated with the FDA when she was involved in this research work. Fessler JA was supported in part by NSF grant IIS 1838179.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The physical phantom CT images and the simulated phantom CT images that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Söderberg M, Gunnarsson M. Automatic exposure control in computed tomography—an evaluation of systems from different manufacturers. *Acta Radiol*. 2010;51(6):625-634.
2. Huck SM, Fung GSK, Parodi K, Stierstorfer K. The z-sbDBA, a new concept for a dynamic sheet-based fluence field modulator in X-ray CT. *Med Phys*. 2020;47(10):4827-4837.
3. Elbakri IA, Fessler JA. Statistical image reconstruction for polyenergetic X-ray computed tomography. *IEEE Trans Med Imaging*. 2002;21(2):89-99.
4. Qi LP, Li Y, Tang L, et al. Evaluation of dose reduction and image quality in chest CT using adaptive statistical iterative reconstruction with the same group of patients. *Br J Radiol*. 2012;85(1018):e906-e911.
5. Chen H, Zhang Y, Kalra MK, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36(12):2524-2535.
6. Chen H, Zhang Y, Zhang W, et al. Low-dose CT via convolutional neural network. *Biomed Opt Express*. 2017;8(2):679-694.
7. Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37(6):1348-1357.
8. Kim B, Han M, Shim H, Baek J. A performance comparison of convolutional neural network-based image denoising methods: the effect of loss functions on low-dose CT images. *Med Phys*. 2019;46(9):3906-3923.
9. MacDougall RD, Zhang Y, Callahan MJ, et al. Improving low-dose pediatric abdominal CT by using convolutional neural networks. *Radiology*. 2019;1(6):e180087.
10. Yang W, Zhang H, Yang J, et al. Improving low-dose CT image using residual convolutional network. *IEEE Access*. 2017;5:24698-24705.
11. Solomon J, Lyu P, Marin D, Samei E. Noise and spatial resolution properties of a commercially available deep learning-based CT reconstruction algorithm. *Med Phys*. 2020;47(9):3961–3971.
12. Lenfant M, Chevallier O, Comby P-O, et al. Deep learning versus iterative reconstruction for CT pulmonary angiography in the emergency setting: improved image quality and reduced radiation dose. *Diagnostics (Basel, Switzerland)*. 2020;10(8):558.
13. Kawashima H, Ichikawa K, Takata T, et al. Performance of clinically available deep learning image reconstruction in computed tomography: a phantom study. *J Med Imaging (Bellingham)*. 2020;7(6):063503.
14. USFDA. *Executive Summary for the Patient Engagement Advisory Committee Meeting: Artificial Intelligence (AI) and Machine Learning (ML) in Medical Devices*. US Food and Drug Administration. Accessed October 22, 2020. https://www.fda.gov/media/142998/download.
15. McCollough CH, Bartley AC, Carter RE, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 Low Dose CT Grand Challenge. *Med Phys*. 2017;44(10):e339-e352.
16. Zeng R, Lin CY, Li Q, Jiang L, Fessler JA, Myers KJ. Generalizability test of a deep learning-based CT image denoising method. *The 6th International Conference on Image Formation in X-Ray Computed Tomography*, 2020.
17. Huber NR, Missert AD, Yu L, Leng S, McCollough CH. Evaluating a convolutional neural network noise reduction method when applied to CT images reconstructed differently than training data. *J Comput Assist Tomogr*. 2021;45(4):544-551.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016; pp. 770-778.
19. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*. 2017;26(7):3142-3155.
20. Ronneberger O, Fischer P, Brox T. U-Net: convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI*. 2015:234-241.

21. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. *Proceedings of the International Conference on Neural Information Processing Systems*. 2014:2672–2680.
22. Zeng R, Divel S, Li Q, Myers KJ. Performance evaluation of deep learning methods applied to CT image reconstruction. *Med Phys*. 2019;46:E162-E162.
23. Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. Towards task-based assessment of CT performance: system and object MTF across different reconstruction algorithms. *Med Phys*. 2012;39(7):4115-4122.
24. Vaishnav JY, Jung WC, Popescu LM, Zeng R, Myers KJ. Objective assessment of image quality and dose reduction in CT iterative reconstruction. *Med Phys*. 2014;41(7):071904.
25. Goldman LW. Principles of CT: multislice CT. *J Nucl Med Technol*. 2008;36(2):57-68.
26. Goodfellow I, Bengio Y, Courville A. *DeepLearning*. The MIT Press; 2016.
27. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.
28. Prabhat KcP, Zeng R, Farhangi MM, Myers KJ. *Deep neural networks-based denoising models for CT imaging and their efficacy*. *Proc SPIE* 2021:11595.
29. Zeng R, Petrick N, Gavrielides MA, Myers KJ. Approximations of noise covariance in multi-slice helical CT scans: impact on lung nodule size estimation. *Phys Med Biol*. 2011;56(19):6223-6242.
30. Li K, Tang J, Chen G-H. Statistical model based iterative reconstruction (MBIR) in clinical CT systems: experimental assessment of noise performance. *Med Phys*. 2014;41(4):041906-041906.

How to cite this article: Zeng R, Lin CY, Li Q, et al. Performance of a deep learning-based CT image denoising method: Generalizability over dose, reconstruction kernel and slice thickness. *Med Phys*. 2022;49:836–853.
https://doi.org/10.1002/mp.15430