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

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

***Abstract*— Deep learning (DL) has been increasingly explored in low-dose CT image denoising. DL products have also been submitted to the FDA for premarket clearance. While having the potential to improve image quality over the filtered back projection method (FBP) and produce images quickly, generalizability of DL approaches is a major concern because the performance of a DL network can depend highly on the training data. In this work we take a residual encoder-decoder convolutional neural network (REDCNN)-based CT denoising method as an example. We investigate the effect of the scan parameters associated with the training data on the performance of this DL-based CT denoising method and identifies the scan parameters that may significantly impact its performance generalizability. This abstract particularly examines these three parameters: reconstruction kernel, dose level and slice thickness. Our preliminary results indicate that the DL network may not generalize well between FBP reconstruction kernels, but is insensitive to slice thickness for slice-wise denoising. The results also suggest that training with mixed dose levels improves denoising performance.**

***Index Terms*—Deep learning, Generalizability test, Low-dose CT denoising**

I. Introduction

CT imaging is widely used in the clinic to assist the diagnosis of abnormalities and to monitor treatment response. It is essential to reduce the x-ray dose to a reasonable level for patient safety while maintaining the CT image quality for accurate decision making. Various approaches have been developed toward this goal through smarter hardware design, such as automatic exposure control, kV optimization and dynamic bowtie filter design, as well as through advanced image reconstruction/ denoising methods, such as statistical and model-based iterative reconstruction (IR) algorithms. Recently, deep learning (DL) methods are attracting high attention, thanks to the growth of big data and increased computation power. DL methods have the potential to improve image quality over FBP and produce image quality comparable to some IR methods (1).

Unlike traditional FBP and (IR) algorithms that were derived based on the imaging physics, DL methods rely on training data that usually contain noisy images and their corresponding high-quality (low noise or better resolution) target to optimize the DL network coefficients to estimate and remove the noise from a noisy input. This data-driven mechanism makes the DL performance less predictable when applying the DL network to process data that has properties that differ from the training data. In CT imaging, these properties include different anatomical regions and different acquisition settings. Considering the varieties of the CT data, generalizability is hence an important aspect in the performance evaluation of DL based denoising methods.

This work aims to investigate the generalizability of DL based CT denoising methods. We implemented an example DL denoising network(2). We categorized the training images in terms of the scan parameters and trained the network separately with each data category. We then evaluated the performances of the DL networks and compared how the performances may differ when tested on the same and different categories of testing data. This abstract reports our initial findings regarding the performance generalizability of DL-based CT denoising methods related to the CT acquisition parameters.

II. Methods

A. Low-dose CT denoising network

Let $\boldsymbol{x} \in R^{m\times n}$ denote a high-noise CT 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 low-noise image $\boldsymbol{y} \in R^{m\times n}$ by minimizing a loss function between $\boldsymbol{x}$ and $\boldsymbol{y}$ based on a giving set of training data. After the network is optimized, a noisy CT image can be passed through the network to produce an image that may have reduced noise.

Various network structures have been explored in the literature for low-dose CT denoising, including fully connected neural network (FCN), residual network (ResNet), UNet and others. In this work, we selected the residual encoder-decoder convolutional neural network (REDCNN) developed by Chen et al.(2) for performing a generalizability test. The loss function for training the denoising network was the mean squared error (MSE) between the DL output and the corresponding low-noise training target.

As illustrated in Fig. 1, REDCNN contains 10 layers of FCNs (5 convolutional and 5 deconvolutional layers) with a rectified linear units (ReLU) activation function following the FCN in each layer. Three shortcuts are added to connect the convolution layers and deconvolution layers to construct a

Corresponding author: Rongping Zeng, rongping.zeng@fda.hhs.gov
Affiliations: Zeng R, Li Q, Lu J and Myers KJ are from the Office of Science and Engineering Laboratories, within the Center for Devices and Radiological Health, US Food and Drug Administration at Silver Spring, Maryland. Lin CY is from the Department of Mathematics and Fessler JA is from the Department of Electrical Engineering and Computer Science in the University of Michigan at Ann Arbor, Michigan. Lin CY was supported in part by the FDA critical path funding and Fessler JA was supported in part by NSF grant IIS 1838179.

residual learning mechanism. More details about the network structure design can be found in the paper by Chen et al.(2).

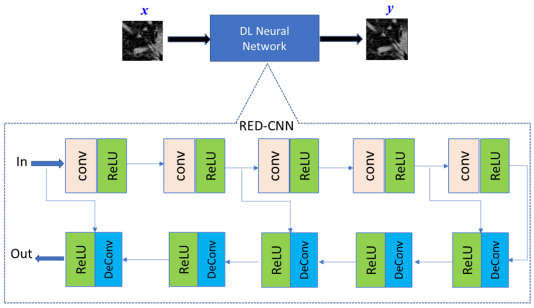

Fig. 1. Illustration of the REDCNN denoising network.

B. Training data categorization

We used the Low-dose Grand Challenge (LDGC) dataset(3) shared by the Mayo Clinic to train the DL denoising network. The LDGC dataset contained ten patient datasets of quarter and full dose scans acquired on a Siemens CT scanner reconstructed with two slice thicknesses (1 mm and 3 mm) and two reconstruction kernels (a sharp kernel D45 and a smooth kernel B30). The corresponding quarter- and full-dose image pairs were treated as training input and training target in the DL training process. Among the ten patient datasets, seven were used for training. A total of 350 slices of size 512x512 were randomly selected from the seven patients and each slice was divided into 55x55 patches excluding the outside of the body. This resulted in about 70,000 training patches.

Note that in the LDGC dataset there was only one reduced dose level (25%) available. To synthesize other dose levels, we extracted the noise by subtracting the quarter dose image from the full dose image and then blended a portion of the noise back into the full dose image. By varying the blending factor from 0.5 to 1.2, we created images corresponding to dose levels ranging about 17% to 80% of the full-dose scan.

The variety of reconstruction thickness, reconstruction kernel and augmented dose levels allowed us to examine the performance generalizability of the DL denoising network for the three parameters. Particularly we categorized the training images into the following six groups based on the combined acquisition parameter values:

- Kernel effect: **Sharp kernel** / 3 mm thickness / 25% dose level; **Smooth kernel** / 3 mm thickness / 25% dose level
- Thickness effect: Smooth kernel / **1 mm thickness** / 25% dose level; Smooth kernel / **3 mm thickness** / 25% dose level
- Dose level effect: Smooth kernel / 3 mm thickness / **25% dose level**; Smooth kernel / 3 mm thickness / **Mixed dose levels of 17%-80%**

With this categorization, we were able to obtain six trained DL networks. For convenience, we name 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. The six networks will be evaluated to check how they may preserve their performance when denoising a different category of testing data relative to their performance on denoising the same category of data as used in training.

C. Generalizability performance evaluation

To evaluate the performance, we considered the following metrics: MSE, modulation transfer function (MTF), noise power spectrum (NPS), and low-contrast detectability (LCD). MSE reflects how well the network performs in terms of minimizing the loss function that the network is designed to do; MTF and NPS are standard performance metrics that characterize the image quality of linear imaging systems; LCD represents a task-based performance metric that augments MTF and NPS to challenge nonlinear smoothing algorithms such as IR and DL methods.

For the MSE measure, the slices from one of the three patients in the LDGC dataset that were not included in the training were used as the test set. The corresponding full-dose images were treated as the references for calculating the MSE. For the MTF, NPS and LCD measures, we simulated 2D phantom images of the CATPHAN600 contrast module (Fig. 2), a uniform water phantom and the CCT189 phantom (Fig. 2), respectively. The simulated fan-beam CT scanner will be described in the next paragraph. The contrast module in CATPHAN600 contains a few disks with varying HU values that allow the measurement of contrast-dependent MTF for non-linear image reconstruction(4). The NPS was estimated by taking the average of the modulus square of the Fourier transform of multiple repeats of the uniform phantom scans. The CCT189 phantom contains four low contrast disks with varying size/HU combinations (3mm/14HU, 5mm/7HU, 7mm/5HU, 10mm/3HU). The task of detecting the low-contrast disks challenges the image reconstruction/denoising methods that usually involve nonlinear smoothing. In this study, the detectability was evaluated using a Laguerre-Gauss channelized Hoteling model observer(5).

The simulated fan-beam CT scanner was set to have distances of 595 mm from the x-ray tube to the isocenter and 1085.6 mm to the detector, same as the Siemens CT scanner used to collect the LDGC dataset. Poisson noise was modeled at the detector and the air photon flux controlled the noise level in the reconstructed images. For FBP reconstruction, we applied two Hann filters of different cutoff frequencies (named Hann1 & Hann2) to be convolved with the sinogram to yield similar MTF50% and MTF10% as the D45 and B30 filters in LDGC (see Table 1). For convenience, we refer to Hann1 and D45 as sharp kernels, Hann2 and B30 as smooth kernels in this abstract. Since the simulated CT scanner was a 2D fan-beam scanner, slice thickness was not a modeled parameter; all the simulated data effectively corresponded to a very thin slice thickness.

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

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

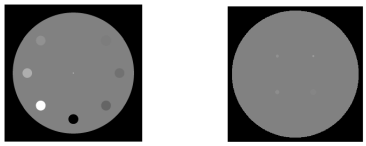

Fig. 2: Left: simulated contrast phantom that mimics the contrast layer in the CATPHAN600 for measuring the MTF. Right: simulated LCD phantom that mimics the CCT189 phantom containing four low-contrast disks for measuring the LCD performance.

III. Results and Discussions

A. MSE

Fig. 3 shows the box plots of the MSE reduction rate of the DL processed image relative to the original noisy image input. Comparison of the kernel effect in the plots clearly shows a significantly larger MSE reduction rate when the test data has the same reconstruction kernel as the training data, indicating that the kernel may be an important factor of the DL performance generalizability. Comparison of the thickness effect shows that the DL network trained with 3mm thickness has slightly better than or similar MSE reduction to the DL network trained with 1mm thickness data, indicating the slice thickness may not be critical. Comparison of the dose effect shows that the DL network trained with a single dose level does not preserve the MSE performance on denoising images at a different dose level but the DL network trained with mixed dose levels can, indicating that training with mixed dose levels is more robust.

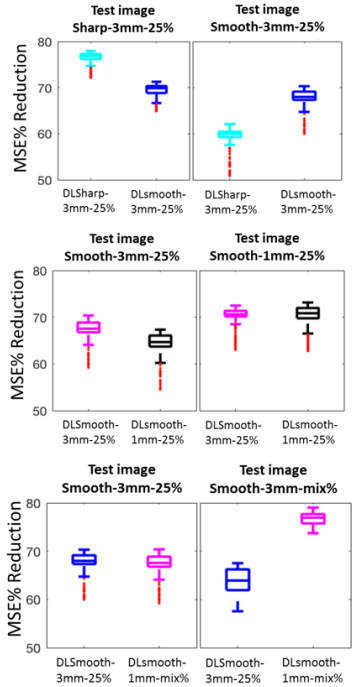

Fig. 3 Comparison of the MSE reduction of the DL networks.

B. MTF

We generated noiseless sinograms of the contrast phantom and reconstructed two images, by FBP of sharp and smooth kernels. We then applied the DL networks to process the two images and measured the MTF at the five contrasts: 990, 340, 200, 120 and 35 HU. The MTF50% value was estimated to represent the image resolution. Fig. 4 plots the MTF50% values vs the HU contrast for the trained DL networks for processing FBP sharp (Fig. 4a) and FBP smooth image (Fig. 4b).

In general, the MTF50% value decreases with the disk HU value, as shown in both plots in Fig. 4. The decreased image resolution with decreased HU is a common characteristic of non-linear image reconstruction or denoising methods. In Fig. 4a, it can be observed that the DL network trained with the sharp kernel data has better image resolution (higher MTF50% value) than the DL network trained smooth kernel, and the DL network trained with mixed dose data has better image resolution than the DL network trained with a single dose data for processing the FBP sharp image. It can also be seen that the DL network trained with 3mm thickness has better image resolution. Similar trends present in Fig 4b for processing the FBP smooth image, although with smaller magnitude of differences among the DL networks.

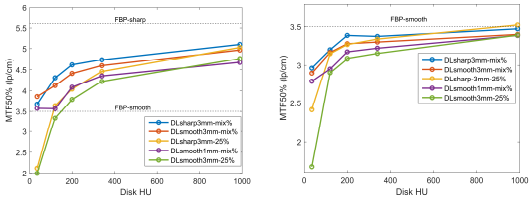

Fig. 4: Contrast-dependent MTF50% curves of the DL networks for processing the FBP sharp (a) and FBP smooth (b) images.

C. NPS

We generated 30 noisy scans of the cylindrical water phantom for the NPS estimation. Each noisy scan was reconstructed by FBP of both sharp and smooth kernels. Based on the MSE and MTF tests, we found that DL network trained with the thicker slice thickness and mixed dose had better performance than those trained with the thin thickness and single dose. Therefore, we decided to only compare the DLsharp-3mm-mixDose and DLsmooth-3mm-mixDose, referred as *DLsharp* and *DLsmooth* thereafter, in the NPS test, as well as in the LCD test later.

Fig.5 shows the 2D NPS images evaluated using the central ROIs of size 64x64. The NPS of the DLsharp appears to have stronger noise magnitude along the northwest diagonal direction, no matter for processing the FBP sharp or smooth images. The NPS of the DLsmooth has a normal round appearance when processing the FBP smooth images but contains much of high-frequency noise when processing the FBP sharp images. By examining the 1D NPS curves (radial binning of the 2D NPS image) shown in Fig. 6, it can be seen that the NPS curve of the DLsmooth has a bump (marked by the blue arrow in the Fig. 6) in the tail that almost matches the magnitude of the NPS tail of the FBP sharp images, indicating that the DLsmooth only suppressed the noise in the low and medium frequency bands but left the high-frequency noise largely untouched. We hypothesize that the smooth kernel data

did not contain high-frequency noise so the DLsmooth did not learn to suppress the high-frequency noise.

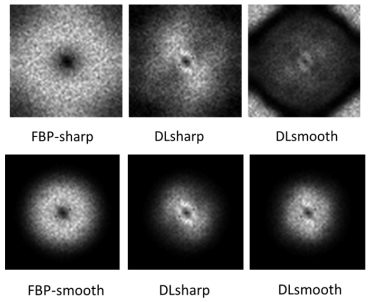

Fig. 5 Top: the NPS images of FBP sharp images and the corresponding DL processed images. Bottom: The NPS images of FBP smooth images and the corresponding DL processed images.

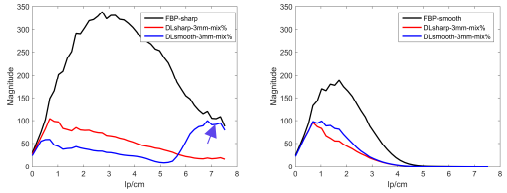

Fig. 6 The 1D NPS curve of the images in Fig. 5. Left: the FBP sharp and the corresponding DL processed images. Right: the FBP smooth and the corresponding DL processed images.

D. LCD

In the LCD test, we simulated 200 noisy scans of the LCD phantom and 100 noisy scans of a uniform phantom at each of the five dose levels (100% to 30%) we selected. The 100% dose level corresponded to an air photon count of 3×10^5. For each low-contrast insert, one signal-present (SP) ROI was extracted from each LCD phantom image and five signal-absent (SA) ROIs were extracted at the vicinity of the insert location from each of the uniform phantom image. As a result, a total of 200 SP and 500 SA ROIs for each insert were created for the evaluation of detectability.

Fig. 7 contains the detectability curves as a function of dose for two of the four inserts. For the two inserts of relatively small sizes (3mm/14HU, 5mm/7HU) the detectability values were very close among the FBP reconstructions and the DL denoising networks, with the AUC ranging from 0.8 to 0.95 in the simulated dose levels, so their detectability curves were not shown in this abstract. The detectability curves are more separated for the two inserts of slightly larger sizes. As clearly shown in the 10mm/3HU insert plot, the smooth kernel boosts the LCD performance in the original FBP reconstruction. The DLsmooth has better LCD performance than the DLsharp on processing both the FBP sharp or smooth images. However, the best LCD performances are achieved by the DLsmooth denoising the FBP sharp image, despite the high-frequency noise shown in the NPS test. We hypothesize that the LCD tasks focused more on the low and medium frequency band and the DLsmooth operated on FBP sharp images achieved an improved balance between image resolution and noise suppression within the signal's frequency band.

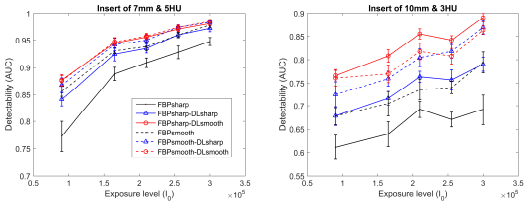

Fig. 7 Detectability curves for the two inserts (7mm&5HU, 10mm&3HU) in the original FBP sharp and FBP smooth images, and denoised FBP images with the DLsharp and DLsmooth.

IV. Conclusions and Future Work

This abstract presented our preliminary work in testing the performance (MSE, MTF, NPS and LCD) generalizability of a DL-based CT denoising method related to three CT acquisition parameters. Our results showed that the DL network was most sensitive to the reconstruction kernel. The DL network trained with thicker slice thickness data appeared to be slightly better than that trained with thinner slice thickness. The DL network trained with mixed dose levels was more robust than that trained with a single dose level. Future work is needed to explore the impact of the other acquisition parameters. Other tasks that can test the preservation of higher frequency information than the LCD task may be necessary for a more complete performance evaluation of the DL denoising networks.

References

1. MacDougall RD, Zhang Y, Callahan MJ, Perez-Rossello J, Breen MA, Johnston PR, Yu H. Improving Low-Dose Pediatric Abdominal CT by Using Convolutional Neural Networks. Radiology: Artificial Intelligence. 2019;1(6):e180087. doi: 10.1148/ryai.2019180087.

2. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. IEEE Transactions on Medical Imaging. 2017;36(12):2524-35. doi: 10.1109/TMI.2017.2715284.

3. McCollough CH, Bartley AC, Carter RE, Chen B, Drees TA, Edwards P, Holmes III DR, Huang AE, Khan F, Leng S, McMillan KL, Michalak GJ, Nunez KM, Yu L, Fletcher JG. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge. Medical Physics. 2017;44(10):e339-e52. doi: 10.1002/mp.12345.

4. 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. Medical Physics. 2012;39(7Part1):4115-22. doi: 10.1118/1.4725171.

5. Vaishnav JY, Jung WC, Popescu LM, Zeng R, Myers KJ. Objective assessment of image quality and dose reduction in CT iterative reconstruction. Medical Physics. 2014;41(7):071904. doi: 10.1118/1.4881148.