On use of augmentation for the DNN-based CT image denoising

Prabhat KC, Kyle J. Myers, M. Mehdi Farhangi, and Rongping Zeng

The Food and Drug Administration, Silver Spring, MD

ABSTRACT

Augmentation strategies have been suggested to overcome issues related to limited training data for the Deep Learning (DL)-based CT image denoising problem. Although augmentation is, indeed, a good machine learning practice, the extent of improvements achieved by DL-based CT denoising solvers through augmentation pipelines remains to be quantified. Accordingly, in this work, we make use of two different deep neural networks (the REDCNN, the DnCNN) to quantify gains in CT image quality through augmentation. The augmentation strategies considered include computer vision inspired strategies (like scaling, rotating, flipping, image-blending) and the CT forward model-based noise (radiation dose) insertion. Likewise, image qualities considered in this work include visual perception- and data fidelity-based global metrics (like the PSNR, the SSIM, the RSE) which are common in the computer vision literature, and CT bench tests (like the NPS, the MTF) and a task-based detectability assessment (the LCD) from the CT imaging literature. Our preliminary results indicate that the DL solvers trained on augmented data show gains in the global metrics, in low-frequency noise texture components and in the MTF values as compared to the ones trained on non-augmented data. However, when the augmented DL-solvers were compared against their low-dose counterparts, their performance - with respect to the noise frequency components, resolution, and detection task - was not all improved, and in some cases even worse.

Keywords: Data Augmentation, Noise Insertion, CT Image Denoising, Deep Learning, Image Quality, Transfer Learning.

1. INTRODUCTION

One of the good machine learning (ML) practices to achieve a generalizable performance by a ML model is to train the model with a large dataset. However, accessibility to sufficiently large data repositories in medical applications – that span across different regions, population demographics, practice of medicine and acquisition protocols – continues to be a challenging prospect. As such, a surrogate data generation methodology called the data augmentation technique is frequently employed to increase the available training data. The prime objective of the data augmentation technique is to ensure that a Deep Learning (DL) method does not overtly train to properties specific only to its training data and rather trains to general properties exhibited by a larger patient population. In the machine learning language, a data augmentation strategy seeks to alleviate issues related to overfitting (when a model performs well on the training dataset but fails to perform on an unseen dataset) and underfitting (when a model fails to perform well on the training dataset). Some of the commonly employed augmentation strategies – borrowed from computer vision-based applications to medical imaging problems – include scaling, rotation, flipping, cropping, and contrast blending. Likewise, there are a broad spectrum of augmentation strategies relevant for medical applications. These strategies can be roughly divided into two categories. These include strategies that replicate variations in (i) image acquisition/practice of medicine and (ii) object/subject/patient being imaged. The category (i) type data augmentations are deduced by making use of physics-based forward models that reflect different medical image acquisition methods. This procedure aids in increasing a given training dataset with respect to covariates such as radiation dose, sparse sampling, scanning parameters etc. For the category (ii) type data augmentation, distribution learning ML models like the Generative Adversarial Networks (GANs) are being explored to synthesize training/testing data that properly

Send correspondence to P. KC
E-mail: prabhat.kc@fda.hhs.gov
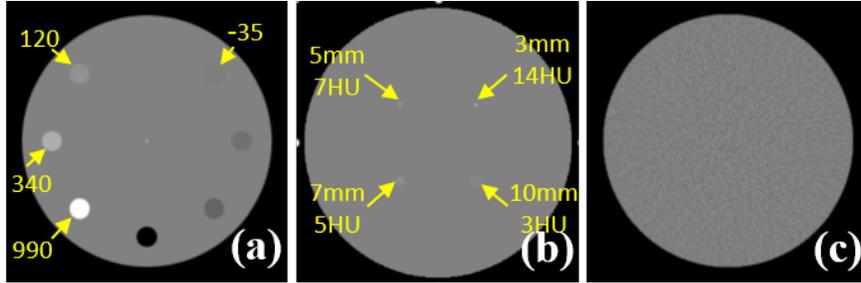
Figure 1: Simulated phantoms that mimic (a) contrast levels in the CATPHAN600 for measuring the MTF, (b) the CCT189 low contrast body phantom for measuring the LCD, and (c) the cylindrical water phantom for the NPS estimation.

replicate variations in image textures found in the patient population (without explicitly performing patient-based acquisitions).

Having said this, the extent of improvements in CT image quality through augmentation strategies for medical applications still remain elusive. In this contribution, we seek to quantify gains of DL-based denoising solvers trained on augmented data. We make use of computer vision-inspired image augmentations and CT physics-based noise insertion as the two data augmentation techniques. The latter augmentation methodology is performed under the framework of transfer learning. Likewise, we make use of two differing DL-based convolutional neural networks (CNNs) for the denoising purpose. These include the feed-forward denoising CNN (DnCNN: that incorporates very deep CNNs to formulate a total of 17 layers) and the Residual Encoder-Decoder CNN (REDCNN: that incorporates encoder-decoder layers to constitutes a total of 10 layers).[1] Finally, we make use of global metrics (such as the Relative Squared Error (RSE), the Peak-Signal to Noise Ratio (PNSR), the Structural Similarity Index (SSIM)), as well as CT bench tests (the contrast-dependent Modulation Transfer Function (MTF), the Noise Power Spectrum (NPS)) and a task-based assessment (the Low-Contrast Detectability (LCD)) to estimate changes in the image qualities of the DL-based denoised solutions trained on the augmented vs non-augmented data.

2. METHOD

2.1 DL framework and architecture

Network weights, θ, of the two DNNs, i.e. the DnCNN and the REDCNN considered in this paper, are estimated by solving the following objective function:

$$\hat{\theta} \leftarrow \arg \min_{\theta} \ell(f(\mathbf{X}; \theta), \mathbf{Y}),\tag{1}$$

where $\ell(\cdot)$ denotes a loss function and $\mathbf{X}, \mathbf{Y} \in \mathbf{R}^{m \times n}$ represent, complimentary, Low-Dose CT (LDCT) and Normal-Dose CT (NDCT) images used to train the two networks. These images are sequestered from the Low-Dose Grand Challenge (LDGC) dataset. The training set comprises 1560 CT images of size 512×512 obtained from 6 different patients. This training data is artificially increased through various procedures (detailed in section 2.2). Likewise, our validation/tuning set consists of 36 CT images randomly pre-selected from the same 6 patients before formulating the training set. Finally, our test sets include 223 CT images obtained from a seventh patient, the CATPHAN600 phantom (fig. 1(a)), the CCT189 phantom (fig. 1(b)) and the cylindrical water phantom (fig. 1(c)). These test sets are not used in any of the network training procedures.

2.2 Dose augmentation framework

The LDCT images in the LDGC dataset correspond to quarter-dose CT (QDCT & $\mathbf{X} = \mathbf{X}_{QD}$) acquisitions. Accordingly, we make use of a CT-physics based forward model[2] to synthesize radiation dose acquisitions corresponding to 95% dose ($\mathbf{X}_{95}$) and 75% dose ($\mathbf{X}_{75}$). Here we note that the normal-dose and quarter-dose projection
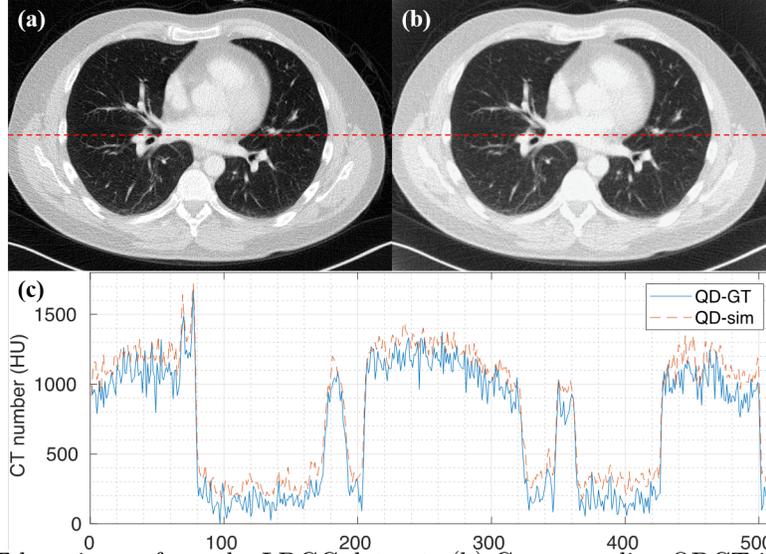
Figure 2: (a) A QDCT lung image from the LDGC dataset. (b) Corresponding QDCT image from our realistic noise insertion model. (c) Line plot comparison of the two images along the dotted red-line. The display window of images in (a) & (b) is (W: 1300 L: -370).

measurements provided in the LDGC were acquired in helical mode. However, in absence of the vendor specific forward model for the helical acquisition (A_{Helical}), we make use of the NDCT images and a CT fan-beam based forward model ($\mathbf{A}_{\mathrm{fan}}$) to synthesize projections in which we insert dose-based noise (η). Mathematically, $\mathbf{X}_{95} = \mathbf{A}_{\mathrm{fan}}^{-1}(\mathbf{A}_{\mathrm{fan}}\mathbf{Y} + \eta_{95\%})$ and $\mathbf{X}_{75} = \mathbf{A}_{\mathrm{fan}}^{-1}(\mathbf{A}_{\mathrm{fan}}\mathbf{Y} + \eta_{75\%})$. The inversion is performed using the Filtered BackProjection (FBP) method. We validated the noise components of the outputs from our realistic dose-based augmentation procedure using the NPS. A representative scan determined from this realistic dose augmentation procedure is provided in fig. 2(b).

Next, we make use of a transfer learning framework to sequentially train the two DNNs on the augmented and non-augmented datasets. This work draws its motivation from the traditional multi-scale based tomographic reconstruction to improve the quality of the restored image. More concretely, rather than performing a straightforward network learning between the quarter- and normal-dose pairs ($\mathbf{X}_{\mathrm{QD}}, \mathbf{Y}$), we sequentially train networks on a series of increasing radiation dose gaps between the input and target pairs. In other words, $\hat{\theta}_{95\text{-ND}}$ is estimated from training pairs ($\mathbf{X}_{95}, \mathbf{Y}$); then $\hat{\theta}_{95\text{-ND}}$ is used as the initializer to estimate $\hat{\theta}_{75\text{-ND}}$ from training pairs ($\mathbf{X}_{75}, \mathbf{Y}$); and finally, $\hat{\theta}_{75\text{-ND}}$ is used as the initializer to estimate $\hat{\theta}_{\mathrm{QD}\text{-ND}}$ from the training pairs ($\mathbf{X}_{\mathrm{QD}}, \mathbf{Y}$).

2.3 Computational optimization

In ref. ,[3] we showed that the performance of a DNN depends on several choices. These choices include patch size (e.g. P-55 for patch 55×55), learning rate (lr), mini-batch size (mi-b), loss function (e.g. MSE, MAE, MSE with L_1 image prior/ total variation (TV) image prior/ weight decay term, MS-SSIM), pre-processing choices (in terms of normalization vs no normalization (*normF*)), and pseudo data augmentation (rotation, scaling, flipping, image-blending). This overall optimization procedure can be compartmentalized into the following three stages:

(i) Stage 1 (*without/ aF*): At this stage, we do not perform any augmentation (*aF*) on the training data. Next, the training data for both the DNNs is normalized to the range $[0, 1]$ (i.e., *unity* normalization). The two DNNs are trained to yield the best values w.r.t the global metrics - PSNR and SSIM - on the tuning set. At the end of this stage, the optimized choices for the two networks include REDCNN: (P-96 | lr:10^{-4} | mi-b:32 | MSE | *unity-aF*), DnCNN: (P-55 | lr:10^{-4} | mi-b:32 | MSE+$\lambda \cdot$L$_1$ | λ:10^{-7}| *unity-aF*).

(ii) Stage 2 (*Pseudo/ PaT*): At this stage, we perform the scaling, rotation, flipping & image-blending augmentations (*PaT*) on the training data. Also, normalization type is tuned to yield the best performance

for the two DNNs w.r.t the HU accuracy. Accordingly, the optimized choices for these two networks include: REDCNN: (P-96 | lr:10^{-5} | mi-b:64 | MSE+λ·TV| λ:10^{-3}| *normF-PaT*), DnCNN: (P-55 | lr:10^{-4} | mi-b:32 | MSE+λ·L$_1$ | λ:10^{-7}| *unity-PaT*).

(iii) Stage 3 (*Realistic/RaT*): At this stage, we train the two DNNs on realistically dose augmented datasets $((\mathbf{X}_{95}, \mathbf{Y}), (\mathbf{X}_{75}, \mathbf{Y}))$ - estimated from the CT noise based forward model - under the framework of transfer learning as detailed in our previous section 2.2. The corresponding choices for the two DNNs include: REDCNN: (P-96 | lr:10^{-5} | mi-b:64 | MSE+λ·TV| λ:10^{-4}| *normF-RaT*), DnCNN: (P-55 | lr:10^{-4} | mi-b:32 | MSE+λ·L$_1$ | λ:10^{-7}| *unity-RaT*).

In summary, stage 1 depicts the scenario where the DL is performed in the absence of any forms of augmentation. Stage 2 and stage 3 depict scenarios where the DL is supplemented by the computer vision-inspired and by the realistic CT noise-based augmentation strategies respectively.

3. RESULTS & DISCUSSIONS

3.1 Global Metrics

Table 1 lists the RSE, PSNR and SSIM values from the two DNNs applied on the QDCT test images, relative to their NDCT counterparts. Columns 3 & 4 in the table indicate that even when the two DNNs are trained *without-augmentation* (*aF*), they show significant improvement in their global metric values as compared to that from the QDCT. There is no further gain in the global metric values for the DnCNN at the stage 2 (*pseudo or PaT*) augmentation and the stage 3 (*realistic or RaT*) augmentation (besides the SSIM value that is best at the *PaT* stage). On the other hand, REDCNN outputs its best global metric-based performance at the *RaT* stage (besides the SSIM value that is best at the *PaT* stage). Overall, this evaluation suggests that the DnCNN (with its 17 convolutional layers that encapsulate batch normalization layers for a faster convergence) does not gain any new CT imaging related information through the augmentation pipeline. On the contrary, the REDCNN (with its encoder-decoder based 10 weight layers that incorporate a rich set of skip connections) shows a gain in its performance through the augmentation pipeline.

Table 1: RSE, PSNR and SSIM values corresponding to the QDCT images and the DnCNN & REDCNN outputs at different augmentation stages.

Network	Metric	Quarter Dose	Stage 1 (aF)	Stage 2 (PaT)	Stage 3 (RaT)
DnCNN	RSE	0.0141 ($\pm$0.003)	**0.0087** (**$\pm$0.005**)	0.0096 ($\pm$0.005)	0.0089 ($\pm$0.005)
	PSNR	32.615 ($\pm$1.075)	**35.129** (**$\pm$2.106**)	34.839 ($\pm$2.236)	35.095 ($\pm$2.35)
	SSIM	0.7483 ($\pm$0.039)	0.9053 ($\pm$0.014)	**0.9062** (**$\pm$0.019**)	0.9039 ($\pm$0.019)
REDCNN	RSE	0.0141 ($\pm$0.003)	0.0089 ($\pm$0.005)	0.0078 ($\pm$0.005)	**0.0075** (**$\pm$0.004**)
	PSNR	32.615 ($\pm$1.075)	35. 097 ($\pm$2. 159)	35.755 ($\pm$2. 248)	**35.868** (**$\pm$2.17**)
	SSIM	0.7483 ($\pm$0.039)	0. 9048 ($\pm$0. 019)	**0.9061** (**$\pm$0.019**)	0.9050 ($\pm$0.019)

3.2 MTF

To evaluate the CT bench tests from the phantoms, first, we generated a noiseless sinogram from the CATPHAN600 (fig. 1 (a)). The sinogram was inverted using the FBP method - paired with a sharp kernel that was used to reconstruct the training data - and the two DNNs were applied on reconstructed CATPHAN600. Subsequently,

MTF50% values at the four contrast disks $(990, 340, 120, -35$ HU) were estimated to represent image resolution (figs. 3 (a,b)). Both the DNNs show significant improvement in their MTF50% values at the *pseudo-augmentation* stage. For the DnCNN, there is a further minor increment in the MTF50% values at the *realistic-augmentation* stage. On the contrary, for the REDCNN, there is a minor decay in the MTF50% values at the *realistic-augmentation* stage as compared to that at the *pseudo-augmentation* stage. Broadly speaking, after the *pseudo-* or *realistic-augmentation* stages, the maximum MTF50% values for the two DNNs – for all four contrast disks – plateaus at or right below their corresponding MTF50% values from the FBP-sharp method.
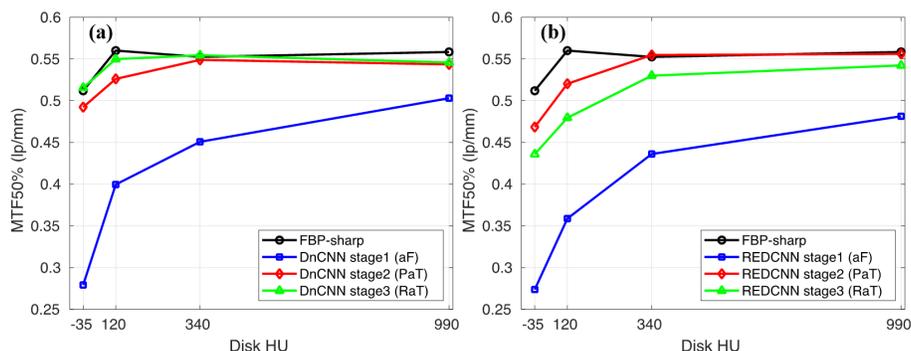
Figure 3: MTF50% plots of the (a) DnCNN and (b) REDCNN applied on the reconstructed CATPHAN600. Each plot illustrates MTF50% values obtained from the FBP method and the different stages of augmentation.

3.3 NPS
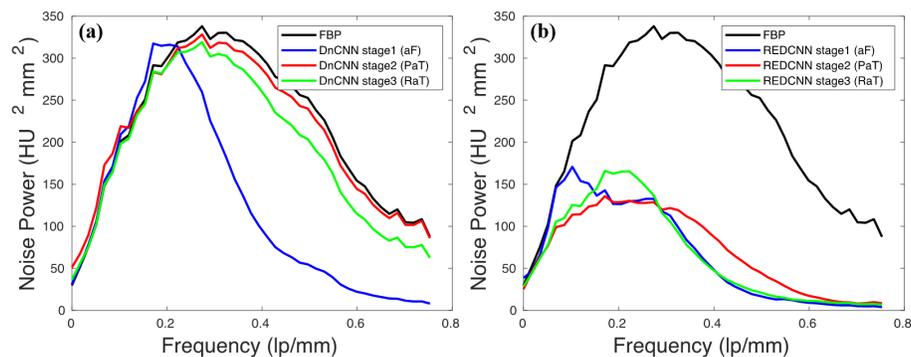
Figure 4: Radial 1D NPS plots evaluated from (a) the DnCNN and (b) the REDCNN applied on the noisy realizations of the reconstructed cylindrical phantom. Each plot includes 1D NPS curves corresponding to the different augmentation strategies.

The NPS was estimated making use of 30 noisy scans/realizations from the cylindrical water phantom that were reconstructed using the FBP method with a sharp kernel. The resulting radial 1D NPS plots for the two DNNs are depicted in fig. 4. In the case of the DnCNN, we notice that for frequency bands above 0.3 lp/mm, there is a significant gain in noise power at the *pseudo-* or *realistic-augmentation* stages as compared to its performance at the *without-augmentation* stage. Also, the DnCNN's performance at the *pseudo-* or *realistic-* stages is similar to that from the FBP method. For the REDCNN, the overall nature of its radial 1D NPS curves - at different augmentation stages - remain mostly similar to one another; thereby suggesting that the REDCNN does not learn to improve noise content information through the augmentation pipeline.

3.4 LCD

For the LCD test, we simulated 200 noisy scans from the CCT189 phantom, and 100 noisy scans from the uniform cylindrical water phantom at each of the five dose levels, i.e. $(30\%, 50\%, 70\%, 85\%, 100\%)$ of incident flux (3×10^5). For each of the four inserts, one signal-present (SP) ROI was extracted from the CCT189 phantom
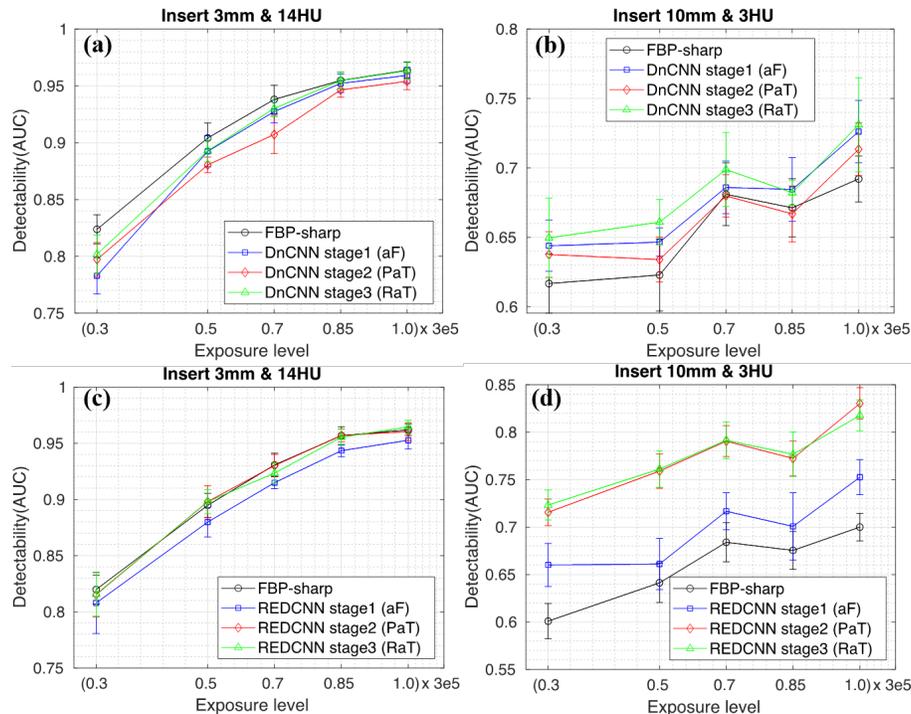
Figure 5: LCD of the DnCNN for (a) insert 3mm, (b) insert 10mm and of the REDCNN for (c) insert 3mm, (d) insert 10mm. Each plot depicts LCD values obtained from the FBP method and the different stages of augmentation.

image and five signal-absent (SA) ROIs were extracted at the vicinity of the insert location from each of the uniform phantom images. As a result, a total of 200 SP and 500 SA ROIs for each low contrast insert were created to evaluate its corresponding detectability. The LCD plots of the two DNNs for the 3 mm & 10 mm inserts are depicted in fig. 5. For the DnCNN (*aF, PaT, RaT*) applied on the 3 mm insert, we do not observe any gain in the detectability of the insert at any of the five dose levels. On the contrary, there is a decay in the detectability performance between 0% to 5% for the DnCNN (*aF, PaT*). We observe similar results for the DnCNN (*aF, PaT, RaT*) applied on 5 mm & 7 mm inserts. However, for the DnCNN (*RaT*) applied on the 10 mm insert, we observe a gain in the LCD values between 0% to 5%, at each dose level, albeit with a large spread in their corresponding error bars. For the REDCNN (*aF, PaT, RaT*) applied on the 3 mm insert, their corresponding detectability values relative to the five dose levels, are around or right below to those obtained from the FBP method. Similar results are seen for the REDCNN (*aF, PaT, RaT*) applied on the 5 mm insert. For the 10 mm insert, we observe a gain in the detectability values ranging from 2% to 5% for the REDCNN (*aF*) and between 11% to 15% for the REDCNN (*PaT, RaT*).

4. CONCLUSIONS

In this contribution, we set out to quantify the improvements in the DL-based CT denoising solvers through augmentation strategies. Correspondingly, for the DnCNN, its very deep network architecture was such that its global metrics-based performance peaks at the *without-augmentation* stage. There is a small drop in its global metrics-based performance at the *pseudo* and *realistic-augmentation* stages. From the LCD viewpoint (that is relevant to diagnostic accuracy), DnCNN yields sub-par performance at *without-augmentation* and *pseudo-augmentation* stages. The gain in its LCD performance was only seen at the *realistic-augmentation* stage, albeit for only large inserts (7 mm & 10 mm) and with a caveat that the detectability of the DnCNN (*RaT*) for low-dose was never as good as that of the FBP at high-dose.

For the REDCNN, its global metric values sequentially increase until the *realistic-augmentation* stage. However, from the LCD viewpoint, the REDCNN's performance saturates at the *pseudo-augmentation* stage.
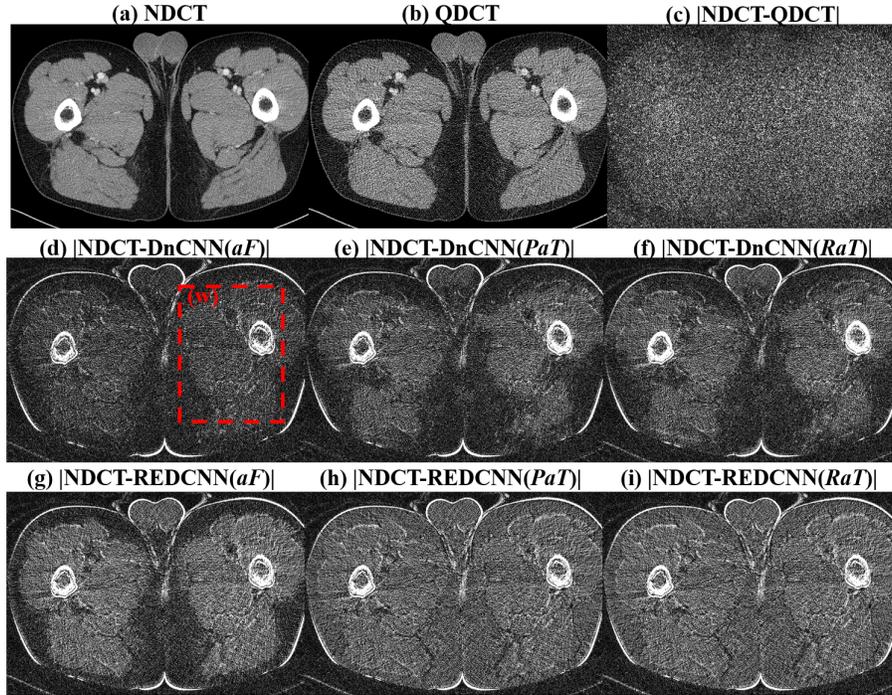
Figure 6: CT images in plots (a) & (b) with the display window of (W:400 L:50) HU. Absolute difference images in plots (c-i) with the display window of [0, 88] HU. Dotted red window (w) is used indicate the central muscular mass with bone & edge contrasts in plots (d-i).

Next, from the resolution viewpoint, for both the DNNs, the *pseudo-augmentation* stage yields significantly higher MTF50% values than that obtained at the *without-augmentation* stage (these values are at or near to the FBP-sharp limit). However, the presence of edge-like structures throughout absolute difference plots (in figs. 6(d-i)) from both the DNNs (for all augmentation stages) points at the limitation of using the digital CATPHAN600 (in particular, using its locally constant regions i.e., disks) to estimate resolution. A proper account of the true nature of the resolution capacity of the two DNNs, relative to different augmentation strategies, will require a further resolution analysis with the aid of a line/wire pattern phantom. Note that the CT scan considered to determine the difference plots is a challenging image for the DL solvers (as it exhibits bone, muscles fibers and fatty regions). Also, the contrast window width of the difference plots is set to properly depict edges in these plots.

Finally, from the NPS viewpoint, it appears that the DnCNN learns not to denoise scans that are devoid of anatomical structures with the sequential augmentation steps (fig. 4(a)). Yet, the visible distinction between the low-contrast outer regions (outside the dotted window (w) in fig. 6) and the high-contrast inner regions (inside the dotted window (w) in fig. 6), and the presence of similar edge-like features (inside (w)) in figs. 6(d-f)) suggest that augmentations (*pseudo* or *realistic*) do not lead to any improvement in low, as well as high, frequency noise components. For the REDCNN, its 1D NPS curves (fig. 4(b)) paired with difference plots (fig. 6(g-i)) indicate that augmentations (*pseudo* or *realistic*) do not lead to any change in high frequency noise components.

In conclusion, when the DL-based denoising solvers were compared against their low-dose counterparts, we did not find any improvement in the solvers' LCD performance, for the 3 mm, 5 mm, & 7 mm, that merits to a statistically significant level. Similarly, the solvers' improvements w.r.t the NPS (evaluated using the uniform phantom) and the MTF (estimated from the contrast disks) do not translate in the noise content information and the resolution of the denoised CT outputs.

In future, we seek to extend this preliminary work with a larger patient data training, and perform more generalizability tests (w.r.t dose levels between QD & ND) and subgroup analyses (by body regions for CT scans vs CT bench testing).

REFERENCES

[1] S. Escalera, S. Ayache, J. Wan, M. Madadi, U. Güçlü, and X. Baró, *Inpainting and Denoising Challenges.* Springer, 2019.

[2] D. Zeng, J. Huang, Z. Bian, S. Niu, H. Zhang, Q. Feng, Z. Liang, and J. Ma, "A simple low-dose x-ray ct simulation from high-dose scan," *IEEE transactions on nuclear science*, vol. 62.

[3] P. KC, R. Zeng, M. M. Farhangi, and K. J. Myers, "Deep neural networks-based denoising models for CT imaging and their efficacy," in *Medical Imaging 2021: Physics of Medical Imaging*, vol. 11595.