Published in final edited form as:

Nat Mach Intell. 2019 June ; 1(6): 269–276. doi:10.1038/s42256-019-0057-9.

Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction

Hongming Shan[1], **Atul Padole**[2], **Fatemeh Homayounieh**[2], **Uwe Kruger**[1], **Ruhani Doda Khera**[2], **Chayanin Nitiwarangkul**[2,3], **Mannudeep K. Kalra**[2,*], **Ge Wang**[1,*]

[1]Biomedical Imaging Center, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA 12180

[2]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA 02114

[3]Division of Diagnostic Radiology, Department of Diagnostic and Therapeutic Radiology, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand 10400

Abstract

Commercial iterative reconstruction techniques help to reduce CT radiation dose but altered image appearance and artifacts limit their adoptability and potential use. Deep learning has been investigated for low-dose CT (LDCT). Here we design a modularized neural network for LDCT and compared it with commercial iterative reconstruction methods from three leading CT vendors. While popular networks are trained for an end-to-end mapping, our network performs an end-to-process mapping so that intermediate denoised images are obtained with associated noise reduction directions towards a final denoised image. The learned workflow allows radiologists-in-the-loop to optimize the denoising depth in a task-specific fashion. Our network was trained with the Mayo LDCT Dataset, and tested on separate chest and abdominal CT exams from Massachusetts General Hospital. The best deep learning reconstructions were systematically compared to the best iterative reconstructions in a double-blinded reader study. This study confirms that our deep learning approach performed either favorably or comparably in terms of noise suppression and structural fidelity, and is much faster than the commercial iterative reconstruction algorithms.

[*]Corresponding authors: MKALRA@mgh.harvard.edu or wangg6@rpi.edu.

Author contributions

G.W. initiated the project. G.W., H.S. and M.K.K. designed the experiments. H.S. and G.W. performed machine learning research. A.P. and F.H. collected CT data and conducted the reader studies. M.K.K, C.N., and R.D.K. were the readers. H.S., G.W., U.K. and M.K.K. analyzed the data. H.S. and G.W. wrote the paper, and M.K.K. and U.K. edited the paper.

Data availability

The training and validation datasets were obtained from the 2016 NIH-AAPM-Mayo Clinic Low-Dose CT Grand Challenge, and are publicly available (http://www.aapm.org/GrandChallenge/LowDoseCT/). The testing datasets from MGH for this study are protected for patient privacy and will be available for data sharing from the corresponding author upon request and after going through an external IRB procedure.

Code availability

The source code is available from the corresponding author Dr. Ge Wang upon request.

Computer vision and image processing are well-known examples to demonstrate the tremendous successes of machine learning, especially deep learning. Both areas take existing images as inputs and produce features of these images as outputs. Conversely, tomographic reconstruction algorithms take measured data, such as line integrals, Fourier/harmonic components, etc., as inputs, and produce images of internal structures as outputs. Recently, deep learning techniques for tomographic image reconstruction have attracted considerable attention.[1–3]

Computed tomography (CT) is a popular imaging modality with applications in biology, medicine, airport security, and other areas.[4] Despite overwhelming evidence of healthcare benefits, the extensive use of CT has raised concerns on potential risk of cancer or genetic damage with x-ray radiation.[5,6] Many clinical indications can be imaged with low-dose CT (LDCT) to minimize the radiation-related risk without significantly compromising the screening or diagnostic performance.[7] In fact, decreasing the CT radiation dose as low as reasonably achievable (the ALARA principle) is the commonly accepted practice, and LDCT has been a hot research topic in the medical imaging field for almost two decades. The reduction of radiation dose, however, increases data noise and can introduce artifacts in reconstructed images, which may adversely affect its diagnostic performance if these problems are not attended to.

To address this challenge, various noise reduction algorithms were proposed for LDCT, which can be categorized into the following categories: 1) sinogram filtration,[8–10] 2) iterative reconstruction,[11–13] and 3) image post-processing.[14–17] Sinogram filtration methods process either raw data or log-transformed data prior to image reconstruction. In the data domain, the well-known noise characteristics can be directly utilized to help the design of sinogram filters. However, the resultant methods often suffer from edge blurring or resolution loss, and sinogram data are usually inaccessible to most researchers. Iterative reconstruction methods optimize an objective function that combines the statistical properties of raw data and prior information on images. Unfortunately, these iterative techniques are time-consuming, require the sinogram data format, involve hyper-parameters that can only be empirically adjusted, and do not offer consistent image quality improvements.[11] Different from these two categories, image post-processing techniques directly process an image that has already been reconstructed from raw data and is publicly available (subject to the patient privacy, which can be addressed by the IRB approval requesting anonymization of images, etc.).

The main motivation of this study is to demonstrate whether deep neural networks perform better than modern commercial iterative reconstruction methods for LDCT and establish a foundation for CT reconstruction algorithms to be empowered by big data and deep learning. For this purpose, our method of choice is a LDCT denoising approach implemented with a novel deep neural network. In addition to the general applicability of the image post-processing strategy, the implication is clear that if the post-processing network performs favorably or comparably to commercially available iterative algorithms, the inclusion of machine learning elements in the sinogram domain, and the reconstruction process in particular, will further increase the merits of the deep learning approach over iterative image reconstruction. As the raw data format was inaccessible to us, we relied on the post-processing approach to compare the machine learning methods for all three major CT vendors on fair grounds. We underline that our intent is to show the superiority of machine learning over iterative reconstruction algorithms implemented by leading industrial CT vendors, instead of comparing CT image quality metrics among these vendors.

With the rapid development of deep learning techniques, convolutional neural networks (CNNs) have recently achieved state-of-the-art results for LDCT image denoising.[15–19] Currently, the deep-learning-based methods only learn the end-to-end mapping from LDCT images to normal dose CT (NDCT) counterparts by minimizing a quantitatively defined loss function. In this context, a conventional loss function may not reflect radiologists' preference well. Here we introduce a Modularized Adaptive Processing (MAP) Neural Network (MAP-NN) for LDCT imaging via progressive denoising. The novelty of this work relates to the application of deep learning for end-to-process denoising mapping with radiologists-in-the-loop so that the LDCT denoising process can be effectively and efficiently guided by domain experts in a task-specific fashion (Fig. 1a). As radiation dose increases from low to high, CT image quality is gradually improved. To a significant degree, this process can be step-wise mimicked through deep learning from a low image quality associated with an order of magnitude less radiation dose to a high image quality of a NDCT reconstruction. A novel aspect of our approach is that it decomposes the overall network into a number of identical network modules. More precisely, each module improves the image quality by a small increment, which can be evaluated by a group of radiologists (Supplementary Fig. 1). This allows determining the optimal number of modules to maximize the diagnostic performance. That is, each network module constitutes a gradual improvement for the denoising task, and all these modules collectively produce a sequence of denoised images (Fig. 1b). Conversely, conventional CNNs are end-to-end denoising networks that produce denoised LDCT images directly. A further benefit of the gradual denoising of LDCT images by the MAP-NN approach is that it can also be taken advantage to reduce the noise level of NDCT images (Fig. 1c) (since the MAP-NN does not only learn to denoise but also encode a noise reduction direction in the module, and this direction allows denoising NDCT images), which is not possible for conventional end-to-end denoising networks. To implement the MAP-NN approach practically, radiologists can evaluate the quality of a denoised image for each CLONE output and, in this way, determine the best image from a sequence of candidate images. This is both cost-effective and user-friendly, and gives a path for deep learning to impact radiology practice proactively. Providing the possibility to use denoised images produced by CLONEs allows using the same MAP-NN for clinical-task-specific as well as image-region-specific applications under the discretion of the radiologist.

Results

MAP-NN versus commercial iterative reconstruction techniques.

To evaluate the performance of the MAP-NN model, 60 patient scans were obtained in total from Massachusetts General Hospital (MGH), half of them underwent routine abdomen CT exams and the rest routine chest CT exams on scanners from vendors A, B, and C respectively and proportionally. All CT exams were acquired on one of the three commercial

CT scanners from GE Healthcare, Philips Healthcare, and Siemens Healthineers, in a randomized order to protect the identities of the scanners. The sinogram data of LDCT were reconstructed with the corresponding commercial iterative reconstruction (IR) techniques and filtered back-projection (FBP) method. Among different IR settings for each vendor, we asked radiologists to choose three clinical used IR methods before the reader study. The three selected IR methods were randomly renamed as IR1, RI2, and IR3 for each vendor. We applied the trained MAP-NN network to LDCT FBP images and produced three DL-denoised images with mapping depth $D = 1, 2, 3$ for all cases, denoted as DL1, DL2, and DL3, respectively. For each patient, two images were selected at representative sites for the reader study, which are susceptible to noise and artifacts (Methods). Each image was evaluated by three radiologists independently in a double-blinded fashion, and scored in terms of two aspects: noise suppression which is to compensate for low-dose induced data noise, and structural fidelity which is directly related to the diagnostic performance. Clinically, the structural fidelity is more general and more important than the noise level, since we cannot have a good structural fidelity if the noise level is too high, and DL methods can suppress the noise greatly at the expense of a major structural loss. In this study, we target the overall comparison between deep learning (DL) and iterative reconstruction (IR) in their respective best forms. Thus, given the noise suppression and structural fidelity scores of two images, we first compared the structural fidelity scores, and then checked their noise suppression scores if the structural fidelity scores were indistinguishable. By the nature of this study, we focus on the overall comparison between the best DL reconstruction and the best IR reconstruction to compare these two competing methodologies (Fig. 2). Let us highlight the comparison in the following two aspects.

Across CT vendors: For vendors A and B, all three readers preferred the best DL reconstruction over the best IR reconstruction for abdominal imaging while the best DL reconstruction was statistically comparable to the best IR reconstruction for the chest scans. For vendor C, DL was statistically comparable to the IR in the abdomen and chest regions. Overall, the conclusion is that the DL method performs better than or comparable to the IR method.

Different body regions: It should be noted that our MAP-NN model was trained on an abdomen dataset based on typical abdomen and chest CT windows. After training the MAP-NN, we evaluated the models for the abdomen and chest regions. In the abdomen cases, the readers rated the DL method better than the IR method for all vendors in all but one case, *i.e.* R1 rated IR comparable to DL on vendor C. Similarly, in the chest cases, all readers rated the DL method better than the IR method on all vendors except that R3 rated IR statistically comparable to DL on vendor C. Overall, the best DL method performs comparably as or better than the best IR method.

Performance in noise suppression and structural fidelity.

Next, we studied the best DL and best IR reconstructions for every selected vendor and each body region in terms of noise suppression and structural fidelity scores (Fig. 3). For that purpose, we show the mean scores and standard deviations for noise suppression and structural fidelity respectively. In Fig. 3a, the DL method achieved a significantly better performance than IR for both scores. In Fig. 3b, the three readers perceived the image quality produced by DL and IR comparable and Fig. 3c–e show that DL achieved a better or comparable performance relative to IR. Finally, Fig. 3f highlights that each reader considered that the DL and IR methods were statistically comparable. As shown in Fig. 2, in some cases, DL and IR gave the same fidelity scores but DL showed a better performance in noise suppression. For the average fidelity score, DL outperformed IR in 12 out of the 18 classes (in 3 of the remaining classes DL and IR performed same), while by average noise suppression, DL was superior in 14 of the 18 classes. Therefore, we concluded that the DL approach performs either favorably or comparably in terms of noise suppression and structural fidelity scores, compared to the IR approach implemented by the three leading CT vendors.

Sample images from three CT vendors.

Fig. 4 presents sample images obtained using the best DL and IR reconstructions respectively from scans on the CT scanners made by the three vendors. The images demonstrate that the DL method for all vendors enabled better noise suppression and structural fidelity than the IR methods. In fact, the IR abdomen CT images (vendors A and B) were deemed unacceptable or limited for noise and fidelity but the images produced by MAP-NN were acceptable for all the three vendors.

Lesion detectability.

Two of the 30 lesions on NDCT (including a sub-centimeter liver lesion and a tiny apical lung nodule) were not seen on LDCT reconstructed with FBP, IR or DL images. The pseudo-lesion (a focus of enhancement) seen on LDCT FBP and IR method was seen on neither DL nor NDCT images. The remaining lesions (28/30) were seen equally well with IR and DL methods. The liver lesion (red arrows) on abdominal CT images from vendor C in Fig. 4 is equally well seen on NDCT, IR and DL reconstructions. Likewise, four lung nodules (green arrows) on chest CT images from vendor B and centrilobular emphysema (blue arrows) on chest CT images from vendor C are seen on all three image sets (NDCT, IR, and DL), with DL images giving slightly better visibility than IR counterparts.

Discussions

The proposed MAP-NN, which has been enhanced by invoking the radiologist-in-the-loop, performs favorably or comparably, relative to the clinically used iterative reconstruction methods implemented by the three leading CT vendors. Once the MAP-NN is trained, the DL-based denoising process is highly efficient (about 100 slices per second per mapping depth) and easy to use in clinical practice, while iterative reconstruction techniques are time-consuming and subject to significant artifacts.

Compared to previously published deep-learning-based denoising networks[14–17,19–23] that learn the denoising mapping from images collected at a specific low-dose setting to the NDCT counterparts, our MAP-NN can be viewed as a significant refinement and a major extension, which learns not only intermediate denoised images through multiple CLONE stages but also the associated noise reduction direction. Then, the number of CLONE modules, also known as the mapping depth, becomes a key parameter, over which the radiologists have the best judgment on the selection of an optimal mapping depth in a task-specific fashion. The MAP-NN with CLONEs permits a cost-effective and user-friendly interface between deep learning and radiologists, enabling the mixed/augmented intelligence beyond what standalone deep learning could achieve. We provide more details about the differences between the conventional denoising model and the proposed progressive denoising model in Supplementary Notes 1 and 2.

For the first time, our MAP-NN systematically demonstrates that the DL approach can provide a similar or better image quality in terms of structural fidelity and noise suppression as compared to the commercial IR methods that are based on image reconstruction directly from raw data. Most importantly for clinical use, the DL approach is computationally much more efficient than IR. Therefore, the DL approach can already effectively compete with the IR solutions, and potentially replace the IR approach. Furthermore, because DL methods can be vendor agnostic, institutions that have CT scanners of various brands and from different vendors can utilize the MAP-NN model to produce similar image appearances, which is not possible for commercial IR techniques. Even though all reconstruction and processing algorithms are commercial products, our post-processing algorithms can be embedded within image viewer software, which is independent from any vendor. Currently, unique changes in image appearance are associated with vendor-specific reconstruction programs. This is an obstacle for large-scale radiomics studies, and could be streamlined using DL techniques in the future.

However, there are some limitations of this study. First, as an overall comparative study, the MAP-NN has not been optimized to either a specific vendor or a particular body region. The collection of more cases in the future will help improve the denoising performance and enhance the statistical significance of the denoising gains over the IR results. Second, LDCT and NDCT slices in a testing set may not be in perfect registration, which can affect the evaluation scores to some degree. Finally, our DL method was selected to be applicable to CT scans from all three vendors from which we cannot have access to raw data. As a result, more powerful DL methods cannot be implemented without the data format. Despite these limitations, our overall conclusion has been encouraging in that DL is either better than or comparable to IR. In collaboration with a vendor, our algorithm can be specifically trained with their data, and achieve an even better performance than what we have described here using our agnostic algorithm. With the availability of raw data, CT denoising can be performed from the sinogram domain to the image space, utilizing all the information for the best denoising results. Clearly, it is the time now for the CT vendors to open the data format, go machine learning, and develop the next generation of CT image reconstruction algorithms in the deep learning framework.

In conclusion, our DL method provides better or comparable image quality compared to commercial IR techniques from three CT vendors, and there is great potential for optimization of DL-based CT reconstruction methods that handle sinogram data directly.

Methods

Training dataset for MAP-NN.

The training dataset we used is an authorized clinical low-dose CT dataset, which was used for *the 2016 NIH-AAPM-Mayo Clinic Low-Dose CT Grand Challenge*. This dataset included normal-dose abdominal CT images that were taken from 10 anonymous patients and the corresponding quarter-dose CT images were simulated by inserting Poisson noise into the projection data for each case to reach a noise level corresponding to 25% of the normal-dose. For network training, 128,000 image patches of size 64 × 64 were randomly selected from the CT images of 5 patients in the Mayo clinical dataset. To validate the performance of the trained networks, 64,000 image patches were randomly selected from the remaining 5 patients.

Testing image acquisition for comparison between deep learning and iterative reconstruction.

To systematically compare the trained deep learning model with the commercial iterative reconstruction, the Human Research Committee of the MGH Institutional Review Board approved this Health Insurance Portability and Accountability Act compliant prospective clinical study. Also, the Human Research Committee of the RPI Institutional Review Board approved the use of these patient data. All included patients have given informed consent prior to their participation in the study. In total, 60 patients datasets were obtained from MGH, half of them undergoing routine abdomen CT exams and the rest routine chest CT exams, on scanners from vendors A, B, and C respectively and proportionally. All CT exams were acquired on one of the three commercial CT scanners from GE Healthcare (Discovery CT750 HD, Waukesha, WI), Philips Healthcare (Brilliance iCT 256, Andover, MA), and Siemens Healthineers (SOMATOM Definition Flash, Germany), in randomized order to protect the identity of the scanners. The LDCT image series were acquired immediately (within 5–10 s) after the acquisition of their normal dose, clinically indicated CT (NDCT) series. The inclusion criteria were adult patients (age > 18 years), who were hemodynamically stable, able to communicate in English, follow instructions, and hold their breath for at least 10 seconds to avoid motion artifacts. Patients undergoing urgent CT, or with known contrast allergy, pregnant women or women planning to become pregnant were excluded from the study. Hemodynamically unstable patients were also excluded. Cross-sectional measurements (anteroposterior and lateral diameters) at the mid-slice location were recorded for all patients. Both the NDCT and LDCT were acquired at identical 100–120 kV, 0.9–1.1 beam pitch, wide detector configuration and 0.5 second gantry rotation time. For the LDCT exams, the tube current was reduced to deliver less than 1 mSv radiation dose to the patients for the LDCT image series (DLP for abdomen LDCT ≤ 65 mGy·cm; DLP for chest CT ≤ 70 mGy·cm). The section thickness for the abdomen and chest CT were the same as used in our standard of care clinical practice (abdomen CT = 5mm; chest CT = 2.5mm). Radiation doses for chest and abdominal CT are summarized in Supplementary Table 1. The inter-vendor differences in radiation doses for NDCT were due to the differences in patient sizes scanned on different CT scanners. For NDCT exams, all scanners automatically adapted the radiation dose by adjusting the tube current with automatic exposure control technique as per our standard of care clinical practice. Thus, in our study LDCT referred to

CT radiation dose less than 1 mSv for both the chest and abdomen CT examinations. It should be noted that our training and testing datasets were from different institutions and geographic regions. We trained our network with the Mayo clinical low-dose CT dataset that contains simulated low-dose and clinical normal-dose CT images. We then applied the trained model to 60 patient scans from MGH, which include mis-matched LDCT and NDCT images. Hence, we only provided the subjective image quality scores from three radiologists instead of quantitative metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) on the testing set.

Image pre-processing.

Hounsfield units (HU) are a dimensionless unit universally used for CT scanners to express CT numbers in a standardized and convenient form. The scale of HU ranges from −1000 HU for air to +~2000 HU for very dense bones. CT windowing is a process in which the grayscale component of the CT image is manipulated in terms of CT numbers, which can change the appearance of the image to highlight particular anatomic structures. Just as the current commercial FBP and iterative reconstruction techniques, it is critical that DL methods are tuned/trained separately on different anatomic body parts to address specific regional/task-oriented needs and avoid artifacts. Indeed, both of our training and testing datasets were prepared in the abdomen and chest windows respectively, since each body region has specific image contrast requirements. The typical windows for soft tissues ranges between [−160, 240] HU for abdomen and between [−1350, 150] HU for chest scans. We normalized the CT images of these two windows to be the range of [0, 1], and trained two different models for abdomen and chest scans respectively.

Commercial iterative reconstruction techniques.

The sinogram data of LDCT were reconstructed with the corresponding commercial iterative reconstruction (IR) techniques and conventional filtered back-projection (FBP) method. Among different IR settings for each vendor, we asked radiologists to choose three clinical used IR methods before the reader study. The three selected IR methods were randomly renamed as IR1, RI2, and IR3 for each vendor. The selected IR methods for GE included adaptive statistical iterative reconstruction (ASIR) at strengths of 50% and 70%, and model-based iterative reconstruction Veo. The selected IR settings for Philips included Idose-level4, IMR-L1-body Routine, and IMR-L1-body-Soft-Tissue (IMR = iterative model reconstruction). For Siemens, the abdomen and chest were reconstructed using different IR methods; the selected IR settings included IRIDIUM (at strengths of 2, 3, and 4) for abdominal imaging and Safire (at strengths of S2, S3, S4) for chest imaging. Thus, image reconstructions were performed using 9 different image reconstructions techniques, and 3 IR methods per patient, from the sinogram data of the LDCT images provided by vendors A, B, and C. The NDCT images for all patients were reconstructed using FBP.

Double-blind subjective image quality evaluation.

All 300 image series (60 patients × 5 series of slices per patient, including one LDCT, one NDCT, and three IR reconstructions) were used in this study. The LDCT FBP served as the input to our DL method. To avoid the vendor information being identified, all slices are saved in the PNG format. For the chest scans, we chose 2 slices from the upper 20% of the chest and the middle of the chest respectively. For the abdominal scans, we chose 2 slices from the mid liver and the pelvis respectively. This is because these are the areas which are susceptible to noise and artifacts. Therefore, we used CT images from 60 patients for testing. For each patient, two images were selected at representative sites for the reasons described above. Three radiologists (M.K.K with 18 years experience, C.N. with 5 years experience, R.D.K. with 4 years experience) independently evaluated all image cases. Subjective image quality evaluation was performed for each NDCT and LDCT cases independently. The LDCT and NDCT labels were provided to radiologists. Then, three DL and three IR images for 60 patients were randomized, and independently reviewed in a double-blinded fashion (Supplementary Note 3). The radiologists were asked to assess each image separately for image noise and structural fidelity using a 4-point scale [1= Unacceptable for diagnostic interpretation; 2= Suboptimal, acceptable for limited diagnostic information only; 3 = Average, acceptable for diagnostic interpretation; and 4 = Better than usual, acceptable for diagnostic interpretation]. The radiologists were also asked to comment on whether any lesions were present. Cohen's kappa statistics for noise and fidelity among three readers on LDCT images shows inter-reader agreement in the range of [0.42, 0.70] as shown in Supplementary Fig. 2.

Network architecture of MAP-NN.

Fig. 1a presents the structure of the proposed Modularized Adaptive Processing Neural Network (MAP-NN) model consisting of multiple Conveying-Link-Oriented Network Encoder-decoders (CLONEs) for LDCT. The MAP-NN network allows progressive denoising operations, which is different from the conventional denoising model that is trained to denoise LDCT and produce a single denoised LDCT image. Formally, the MAP-NN can be formulated as follows:

$$\boldsymbol{I}_{\mathrm{den}} = g^{T}(\boldsymbol{I}_{\mathrm{LD}}) = \underbrace{(g \circ g \circ \cdots \circ)g}_{T\,=\,\#g}(\boldsymbol{I}_{\mathrm{LD}}) \approx \boldsymbol{I}_{\mathrm{ND}} \tag{1}$$

where $\boldsymbol{I}_{\mathrm{den}}$, $\boldsymbol{I}_{\mathrm{LD}}$, and $\boldsymbol{I}_{\mathrm{ND}}$ denote a denoised image, an LDCT FBP image, and an NDCT FBP image, respectively. The operator $\bigcirc$ denotes a functional composition operation, g denotes an CLONE module, g^t denotes the t-fold product of the CLONE g, and the number of CLONEs for training is denoted by T. The parameters of all the CLONEs are shared. With the Mayo LDCT Challenge dataset, we trained the MAP-NN model for $T = 5$. The progressively denoised images obtained from the trained MAP-NN ($T = 5$) are shown in Fig. 1b,c.

Actually, the module g can be any existing denoising network, such as a fully-connected convolutional networks,[14,17,20,21] a convolutional encoder-decoder network with skip connections,[15,24] a convolutional encoder-decoder with conveying-paths,[19,25,26] stacked denoising autoencoder,[22] and their 3D variants.[16,19,23] In this study the Conveying-Link-Oriented Network Encoder-decoder (CLONE) is an extension to our earlier direct LDCT denoising network, Conveying-Path-based Convolutional Encoder-decoder (CPCE),[19] by coupling with skip connection and output clipping and modularizing into a progressive denoising model. A main merit of our selected CLONE is that the conveying links in the CPCE render it compact and effective. More specifically, the CPCE denoising network has 4 convolutional layers, each of which has 32 filters of size 3 × 3, followed by 4 deconvolutional layers also with 32 filters of size 3 × 3, except for the final layer that has only 1 filter. The filter stride is set to 1 for all convolutional and deconvolutional layers. The conveying path, originally introduced in the U-net[27] for biomedical image segmentation, copies the early feature-maps and reuses them as the input to a later layer of the same feature-map size in a network. This mechanics preserves details of the high-resolution features. Our CLONE has three conveying paths, copying the output of an early convolutional layer and reusing it as the input to a later deconvolutional layer of the same feature-map size. To reduce the computational cost, one convolutional layer with 32 filters of size 1 × 1 is used after every conveying path, reducing the number of feature-maps from 64 to 32. Each convolutional or deconvolutional layer is followed by a rectified linear unit (ReLU). Training a deep progressive denoising network with the above CPCE module, however, can be very difficult. Among many issues, a major problem is the exploding/vanishing gradients.[28] Another problem is that storing an exact copy of information of many layers is not easy. In LDCT denoising, NDCT images are quite similar to LDCT counterparts, and processing modules need to keep the exact copy of input images for many modules. To address these problems, we improved the CPCE module by incorporating the output clipping and the residual skip connection, forming the CLONE unit for progressive denoising model. Because the output of each CLONE is an image, which can be viewed as the “bottleneck” in the progressive process. Therefore, we clip the output into a range of [0, 1], the same as that of the input range, to prevent gradients from becoming extreme. To avoid vanishing gradients, we use a residual skip connection from the input to the output of each module.[29] Hence, each module infers the noise distribution instead of the whole image, compressing the output space and avoiding the use of the exact copy of the input image.

The number of CLONEs, T, is a key parameter in our MAP-NN, the effect of increasing T can be summarized in the following five aspects:

1. the model capacity of MAP-NN will increase (without introducing new parameters);
2. the training difficulty will increase due to gradient vanishing;
3. each CLONE will remove less noise from its input image;
4. more denoised images can be provided for the deliberation of radiologists; and
5. the difference in two consecutively denoised images (outputs from the k-th and the $(k + 1)$-th CLONEs) reduces, which allows finer tuning of the resultant image quality at the expense of an increased workload for radiologists.

In the experiment, the number of CLONES (T) was selected to be 5 for the training data set (Supplementary Fig. 3), as the performance of the MAP-NN (evaluated for the output of the last CLONE) was deemed optimal by radiologists. This decision was made in comparison to various numbers of CLONES, ranging from 3 to 7. In addition to that, the radiologists also evaluated the sequence of denoised images (each CLONE produced in series) and found that

- the difference between two consecutive images drastically increased as T reduces from 5 and

- the gain is insignificant for $T = 6$ or more.

It should be noted that the LDCT images in the training set were realistically simulated by the Mayo Clinic for the AAPM low-dose CT contest to be particularly noisy, which allowed evaluating the MAP-NN rigorously. In contrast, the testing data set, which are images from real clinical scanners, were less noisy. The radiologists conducted the analysis and concluded that the output of the 3rd CLONE is satisfactory, whilst the outputs of the 4th and 5th CLONEs lost some structural fidelity. Thus, the first three denoised images from each CLONEs were used for the reader study, denoted as DL1, DL2, and DL3, respectively.

Composite loss functions.

The composite function for optimizing the network includes three components: adversarial loss, mean-squared error, and edge incoherence. The adversarial loss encourages the generator network to produce samples that are indistinguishable from the NDCT images, which refers to the loss function of the generator. The adversarial loss this study used is defined within the Wasserstein generative adversarial network framework with gradient penalty (WGAN-GP):[30]

$$\min_{\boldsymbol{\theta}_g} \mathscr{L}_a = \mathbb{E}_{\boldsymbol{I}_{\mathrm{LD}}}\left[D\left(g^T(\boldsymbol{I}_{\mathrm{LD}})\right)\right], \tag{2}$$

where network D is called discriminator aiming at distinguishing generated data from ground-truth samples and $\boldsymbol{\theta}_g$ denotes the parameters in network g. The literature has suggested to iteratively optimize the generator g^T once and discriminator D four times,[30–32] which we used in this project. In the experiments, the discriminator D was the same one used in the recent works,[17,19] which has 6 convolutional layers with 64, 64, 128, 128, 256, and 256 filters of size 3 × 3, followed by 2 fully-connected layers of sizes 1024 and 1, respectively. Each convolutional layer is followed by a leaky ReLU, which has a negative slope of 0.2 when the unit is saturated and not active. A unit filter stride is used for oddly indexed convolutional layers, and this stride is doubled for evenly numbered layers (Supplementary Fig. 3).

The mean-squared error (MSE) measures the difference between the output and NDCT images, which would reduce the noise in the input LDCT image. Formally, the MSE is defined as follows:

$$\min_{\boldsymbol{\theta}_g} \mathscr{L}_m = \mathbb{E}_{(\boldsymbol{I}_{\mathrm{LD}}, \boldsymbol{I}_{\mathrm{ND}})}\left\|\boldsymbol{I}_{\mathrm{ND}} - g^T(\boldsymbol{I}_{\mathrm{LD}})\right\|^2. \tag{3}$$

Edge incoherence measures the difference between the Sobel filtrations of real and estimated images as

$$\min_{\boldsymbol{\theta}_g} \mathscr{L}_e = \mathbb{E}_{(\boldsymbol{I}_{\mathrm{LD}}, \boldsymbol{I}_{\mathrm{ND}})}\left\|SF(\boldsymbol{I}_{\mathrm{ND}}) - SF\left(g^T(\boldsymbol{I}_{\mathrm{LD}})\right)\right\|^2 \tag{4}$$

where SF denotes the Sobel filteration corresponding to gradient vector at each point in the image. This filtration is based on convolving the image with a small, separable, and integervalued filter in the horizontal and vertical directions. As a result, the gradient approximation that it produces is relatively crude, in particular for high-frequency variations in the image, which could enhance the edge information in the denoised image.

The final objective function for minimizing MAP-NN is defined as

$$\min_{\theta_g} \mathscr{L} = \mathscr{L}_a + \lambda_m \mathscr{L}_m + \lambda_e \mathscr{L}_e, \tag{5}$$

which encourages the generated denoised image to preserve more texture information, reduce the noise, and enhance the edge.

Network training

The network was optimized using the Adam optimization method[33] with a min-batch of 128 image patches for each iteration. The learning rate was set to 1.0 × 10^{-4} with two exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the moment estimates. The learning rate was adjusted by $1/\sqrt{t}$ decay after every epoch. We tuned the hyper-parameters ($\lambda_m = \lambda_e = 50$) in collaboration with the radiologists in the training stage. Our computational platform consists of four GeForce GTX 1080 TI GPUs with 96GB RAM. The network was trained in 80 epochs within 24 hours on a NVIDIA 1080Ti GPU. We modified the WGAN-GP open source (https://github.com/igul222/improved_wgan_training) for this work, in which networks were implemented with the TensorFlow Library.[34] Two validation curves are shown in Supplementary Fig. 4 for abdomen and chest CT windows, respectively.

Significant test for overall comparison.

We focus on the comparison between the best DL reconstruction and the best IR reconstruction to compare these two competing methodologies. Therefore, we considered two hypothesis tests, involving

- **DL > IR**: The best DL reconstruction outperforms the best IR reconstruction in terms of the structural fidelity; or the best DL reconstruction outperforms the best IR reconstruction in terms of noise suppression score when their structural fidelity scores are indistinguishable;
- **DL = IR**: There are no significant differences between the best DL reconstruction and the best IR reconstruction in terms of structural fidelity and noise suppression;
- **DL < IR**: This is the opposite case of the first hypothesis.

The statistical statistical hypothesis tests were applied in sequence, where the first hypothesis test is:

$$H_0^{(1)}: \mathbf{DL} = \mathbf{IR} \quad H_1^{(1)}: \mathbf{DL} > \mathbf{IR},$$

If the experimental evidence did not allow us to reject $H_0^{(1)}$, we applied the second hypothesis test:

$$H_0^{(2)}:\mathbf{DL} = \mathbf{IR} \quad H_1^{(2)}:\mathbf{DL} < \mathbf{IR}.$$

These two hypotheses were based on the counts #(preferred best DL), #(no preference between best DL and best IR), and #(preferred best IR) among 20 observations per vendor per body region per reader, where # refers to the number of observations in each class. The resultant p-values by the sign test[35] are reported in Supplementary Table 2.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would thank Drs. Michael Vannier and Bruno De Man for helpful discussions, and NVIDIA Corporation for the donation of GPU used for this research.

References

[1]. Wang G A perspective on deep imaging. IEEE Access 4, 8914–8924 (2016).

[2]. Wang G, Ye JC, Mueller K & Fessler JA Image reconstruction is a new frontier of machine learning. IEEE Trans. Med. Imaging 37, 1289–1296 (2018). [PubMed: 29870359]

[3]. Zhu B, Liu JZ, Cauley SF, Rosen BR & Rosen MS Image reconstruction by domain-transform manifold learning. Nature 555, 487 (2018). [PubMed: 29565357]

[4]. Brenner DJ & Hall EJ Computed tomography—an increasing source of radiation exposure. New Engl. J. Med 357, 2277–2284 (2007). [PubMed: 18046031]

[5]. de González AB et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch. Intern. Med 169, 2071–2077 (2009). [PubMed: 20008689]

[6]. Smith-Bindman R et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch. Intern. Med 169, 2078–2086 (2009). [PubMed: 20008690]

[7]. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. New Engl. J. Med 2011, 395–409 (2011).

[8]. Wang J, Lu H, Li T & Liang Z Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters. In Proc. of SPIE, vol. 5747, 2059 (2005).

[9]. Manduca A et al. Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. Med. Phys 36, 4911–4919 (2009). [PubMed: 19994500]

[10]. Wang J, Li T, Lu H & Liang Z Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography. IEEE Trans. Med. Imaging 25, 1272–1283 (2006). [PubMed: 17024831]

[11]. Geyer LL et al. State of the art: Iterative CT reconstruction techniques. Radiology 276, 339–357 (2015). [PubMed: 26203706]

[12]. Willemink MJ et al. Iterative reconstruction techniques for computed tomography part 2: Initial results in dose reduction and image quality. Eur. Radiol 23, 1632–1642 (2013). [PubMed: 23322411]

[13]. Zheng X, Ravishankar S, Long Y & Fessler JA PWLS-ULTRA: An efficient clustering and learning-based approach for low-dose 3D CT image reconstruction. IEEE Trans. Med. Imaging (2018).

[14]. Chen H et al. Low-dose CT via convolutional neural network. Biomed. Opt. Express 8, 679–694 (2017). [PubMed: 28270976]

[15]. Chen H et al. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging 36, 2524–2535 (2017). [PubMed: 28622671]

[16]. Wolterink JM, Leiner T, Viergever MA & Išgum I Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans. Med. Imaging 36, 2536–2545 (2017). [PubMed: 28574346]

[17]. Yang Q et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging 37, 1348–1357 (2018). [PubMed: 29870364]

[18]. Kang E, Chang W, Yoo J & Ye JC Deep convolutional framelet denosing for low-dose CT via wavelet residual network. IEEE Trans. Med. Imaging 37, 1358–1369 (2018). [PubMed: 29870365]

[19]. Shan H et al. 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. IEEE Trans. Med. Imaging 37, 1522–1534 (2018). [PubMed: 29870379]

[20]. Choi K, Kim SW & Lim JS Real-time image reconstruction for low-dose CT using deep convolutional generative adversarial networks (GANs) In Medical Imaging 2018: Physics of Medical Imaging, vol. 10573, 1057332 (International Society for Optics and Photonics, Houston, TX, USA, 2018).

[21]. Kim B, Han M, Shim H & Baek J Performance comparison of convolutional neural network based denoising in low dose CT images for various loss functions In Medical Imaging 2019: Physics of Medical Imaging, vol. 10948, 1094849 (International Society for Optics and Photonics, San Diego, CA, USA, 2019).

[22]. Liu Y & Zhang Y Low-dose CT restoration via stacked sparse denoising autoencoders. Neurocomputing 284, 80–89 (2018).

[23]. You C et al. Structurally-sensitive multi-scale deep neural network for low-dose CT denoising. IEEE Access 6, 41839–41855 (2018). [PubMed: 30906683]

[24]. Hu Z et al. Artifact correction in low-dose dental CT imaging using Wasserstein generative adversarial networks. Med. Phys (2019).

[25]. Gholizadeh-Ansari M, Alirezaie J & Babyn P Deep learning for low-dose CT denoising using perceptual loss and edge detection layer. arXiv Preprint at https://arxiv.org/abs/1902.10127 (2019).

[26]. Yi X & Babyn P Sharpness-aware low-dose CT denoising using conditional generative adversarial network. J. Digit. Imaging 1–15 (2018). [PubMed: 28744581]

[27]. Ronneberger O, Fischer P & Brox T U-Net: Convolutional networks for biomedical image segmentation In Med. Image Comput. Comput. Assist. Interv. (MICCAI), 234–241 (Springer, Munich, Germany, 2015).

[28]. Pascanu R, Mikolov T & Bengio Y On the difficulty of training recurrent neural networks In Proc. of Int. Conf. on Machine Learning (ICML), 1310–1318 (JMLR, Atlanta, GA, USA, 2013).

[29]. Tai Y, Yang J & Liu X Image super-resolution via deep recursive residual network In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (IEEE, Honolulu, HI, USA, 2017).

[30]. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V & Courville AC Improved training of Wasserstein GANs In Advances in Neural Information Processing Systems (NIPS), 5769–5779 (Curran Associates, Long Beach, CA, USA, 2017).

[31]. Goodfellow I et al. Generative adversarial nets In Advances in Neural Information Processing Systems (NIPS), 2672–2680 (Curran Associates, Montreal, Quebec, Canada, 2014).

[32]. Arjovsky M, Chintala S & Bottou L Wasserstein generative adversarial networks In Prof. of Int. Conf. on Machine Learning (ICML), 214–223 (JMLR, Sydney, Australia, 2017).

[33]. Kingma D & Ba J Adam: A method for stochastic optimization. In Proc. of Int. Conf. on Learning Representations (ICLP) (San Diego, CA, USA, 2015).

[34]. Abadi M et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv Preprint at https://arxiv.org/abs/1603.04467 (2016).

[35]. Conover WJ & Conover WJ Practical Nonparametric Statistics Ch. 3 (Wiley New York, 1980).

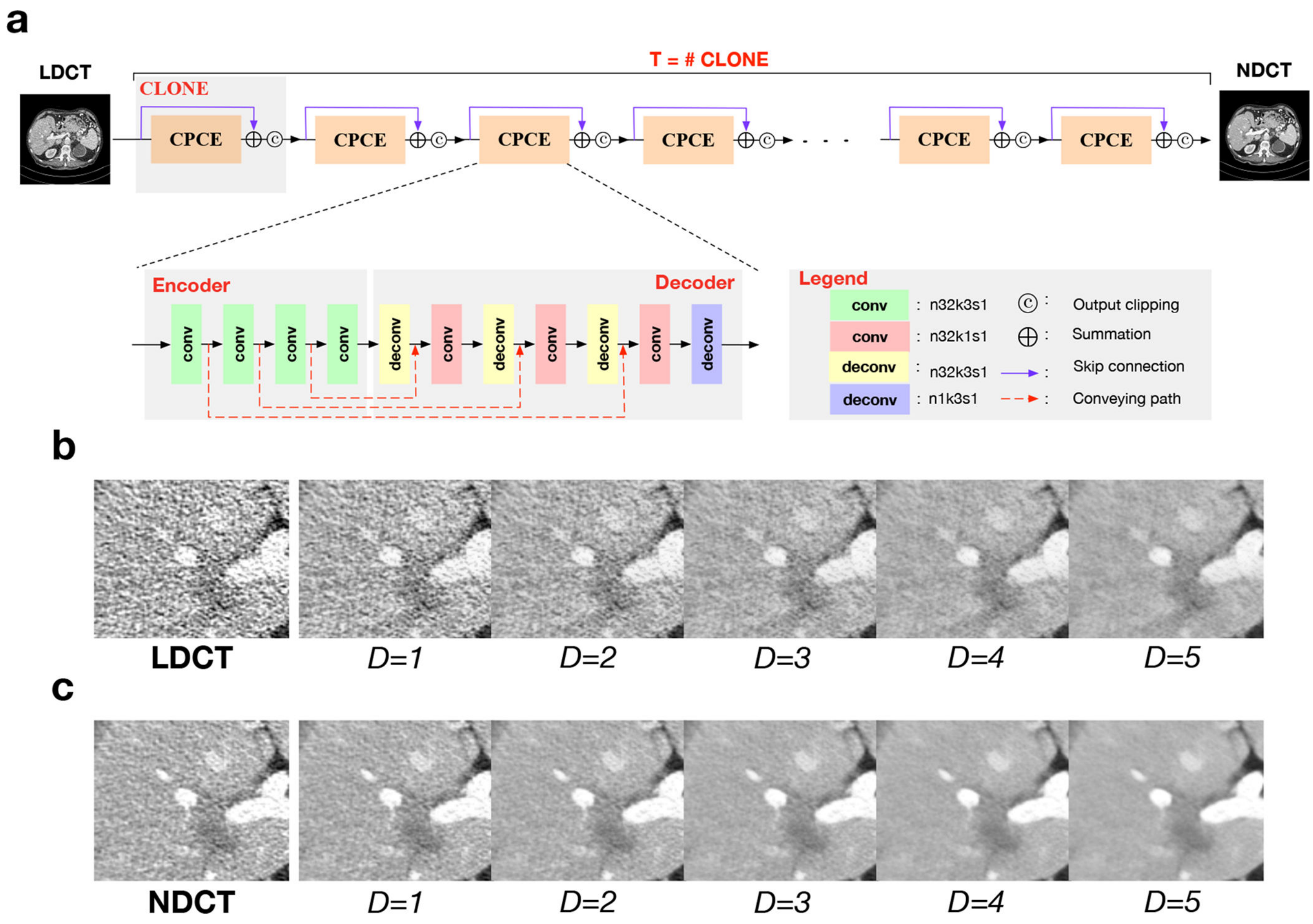

Figure 1: Proposed Modularized Adaptive Processing (MAP) Neural Network (MAP-NN) for LDCT denoising and progressively denoised images on the LDCT and NDCT images with *D* being the mapping depth.

a, Each processing module is a Conveying-Link-Oriented Network Encoder-decoder (CLONE) with skip connection linking input to output, a Conveying-Path-based Convolutional Encoder-decoder (CPCE) network learning the residual between its input and output, a summation operation adding the input and residual, and an output clipping avoiding an exploding gradient. *n*32*k*3*s*1 indicates 32 filters of kernel size 3 × 3 with a stride of 1, each (de)convolutional layer is followed by a ReLU; **b**, Applying the trained MAP-NN to LDCT images from the Mayo dataset confirms that the MAP-NN can learn the denoising direction from LDCT images to NDCT counterparts, and produce intermediate denoised results that improve the image quality progressively; **c**, Applying the trained MAP-NN to NDCT images from the Mayo dataset shows that the MAP-NN model can be used for images acquired at different dose levels to improve the image quality to various degrees. The optimal mapping depth *D* can be visually judged by a group of radiologists; in other words, with radiologists-in-the-loop the denoised image quality can be optimized in a task-specific fashion.

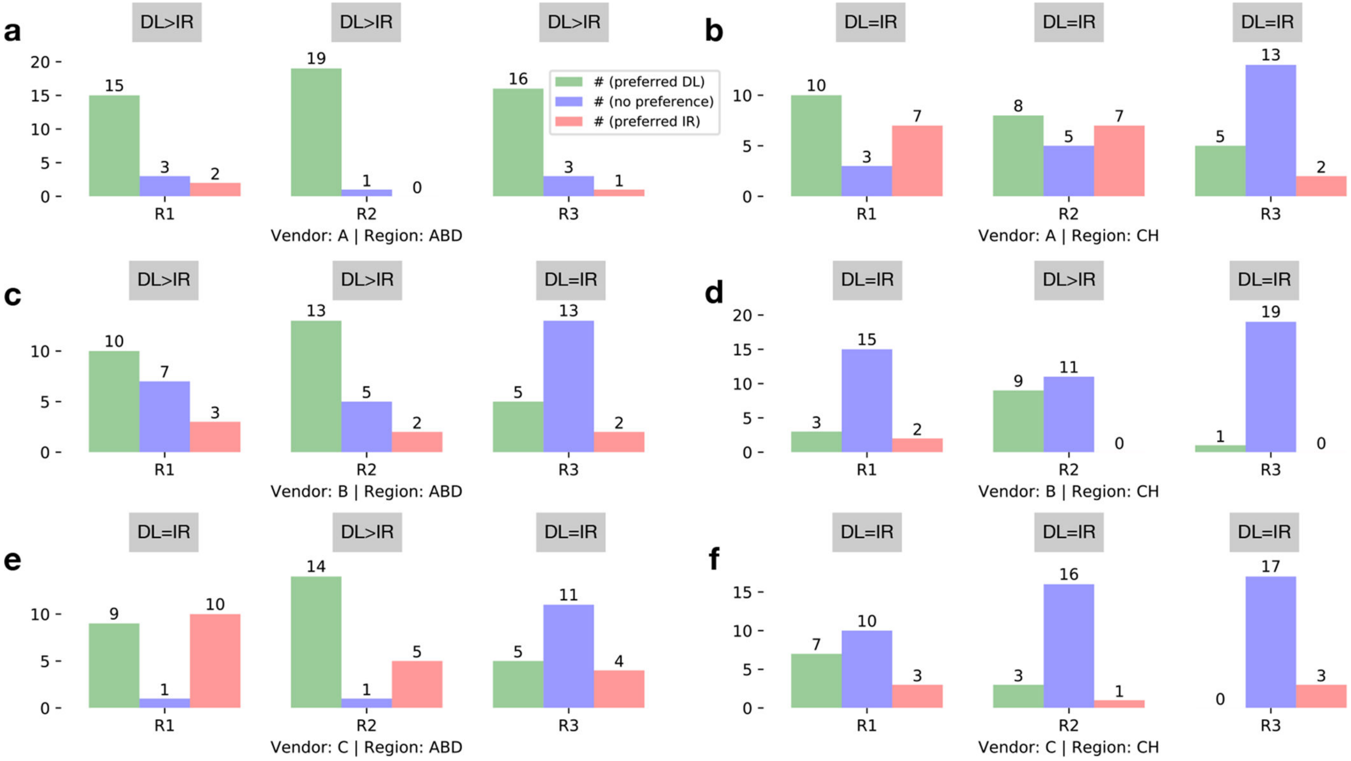

Figure 2: Comparison between the best deep learning (DL) reconstruction and the best iterative reconstruction (IR) for abdomen (ABD) and chest (CH) regions across three major vendors (A, B, and C) and three readers (R1, R2, and R3).

a–f, The histogram showing the number of cases per class (#(preferred DL), #(no preference), and #(preferred IR)) in 20 cases on abdomen from vendor A, chest from vendor A, abdomen from vendor B, chest from vendor B, abdomen from vendor C, and chest from vendor C. The text in the gray box above each plot gives the significant results evaluated by the sign test at 5% significant level for each reader.

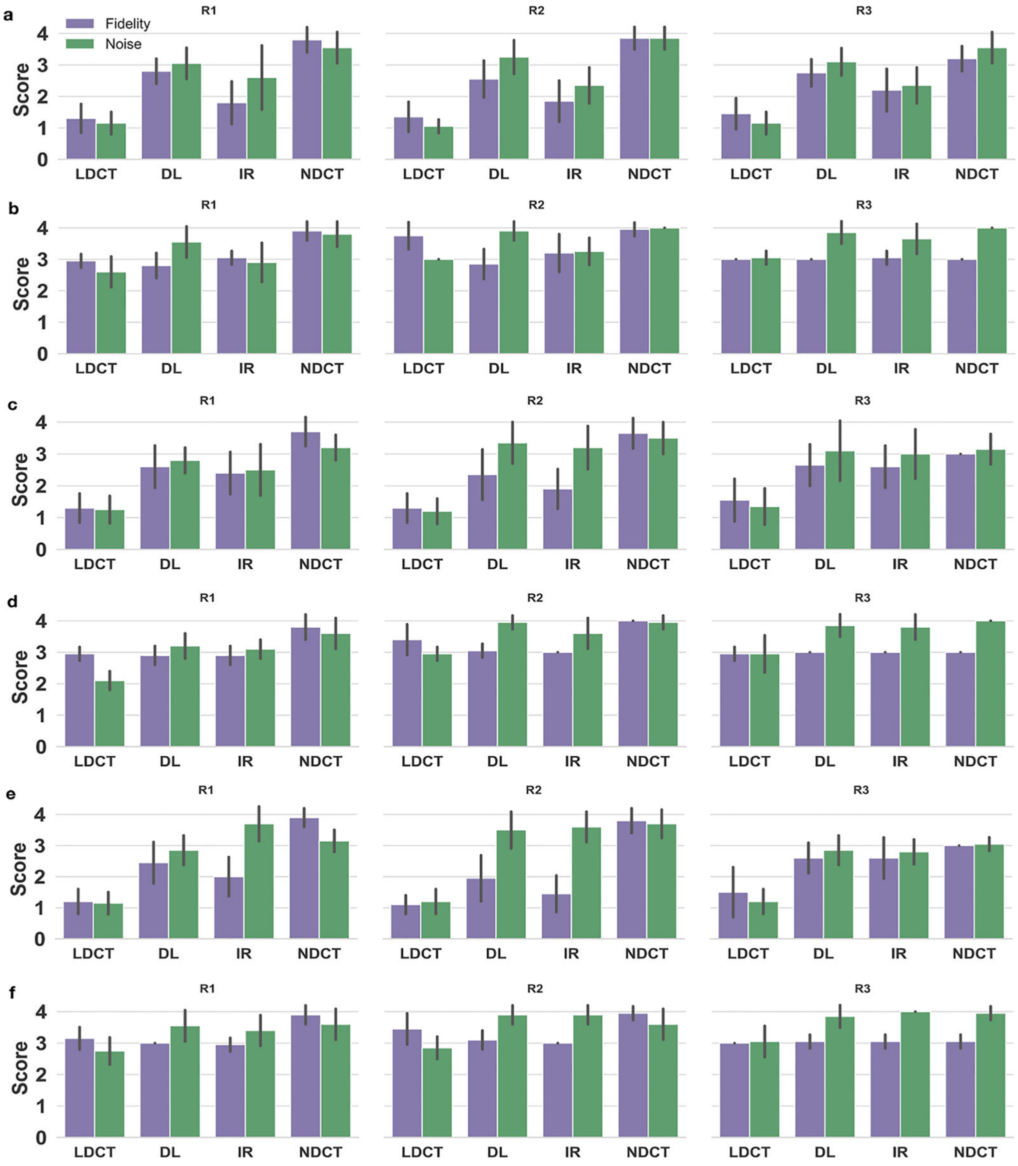

Figure 3: Individual metric comparison between the best DL and best IR reconstructions keyed to body regions (Abdomen and Chest) and CT vendors (A, B, and C).

a–f, Average score and standard deviation, for a sample size of 20, of the noise suppression and structural fidelity scores on abdomen scans from vendor A, chest scans from vendor A, abdomen from vendor B, chest from vendor B, abdomen from vendor C, and chest from vendor C. The performance metrics are fidelity scores and noise suppression for LDCT, best DL, best IR, and NDCT. In terms of average fidelity scores, DL outperformed IR in 12 of 18 classes (in 3 of the remaining classes DL and IR performed same), while by average noise suppression, DL was superior to IR in 14 of 18 classes.

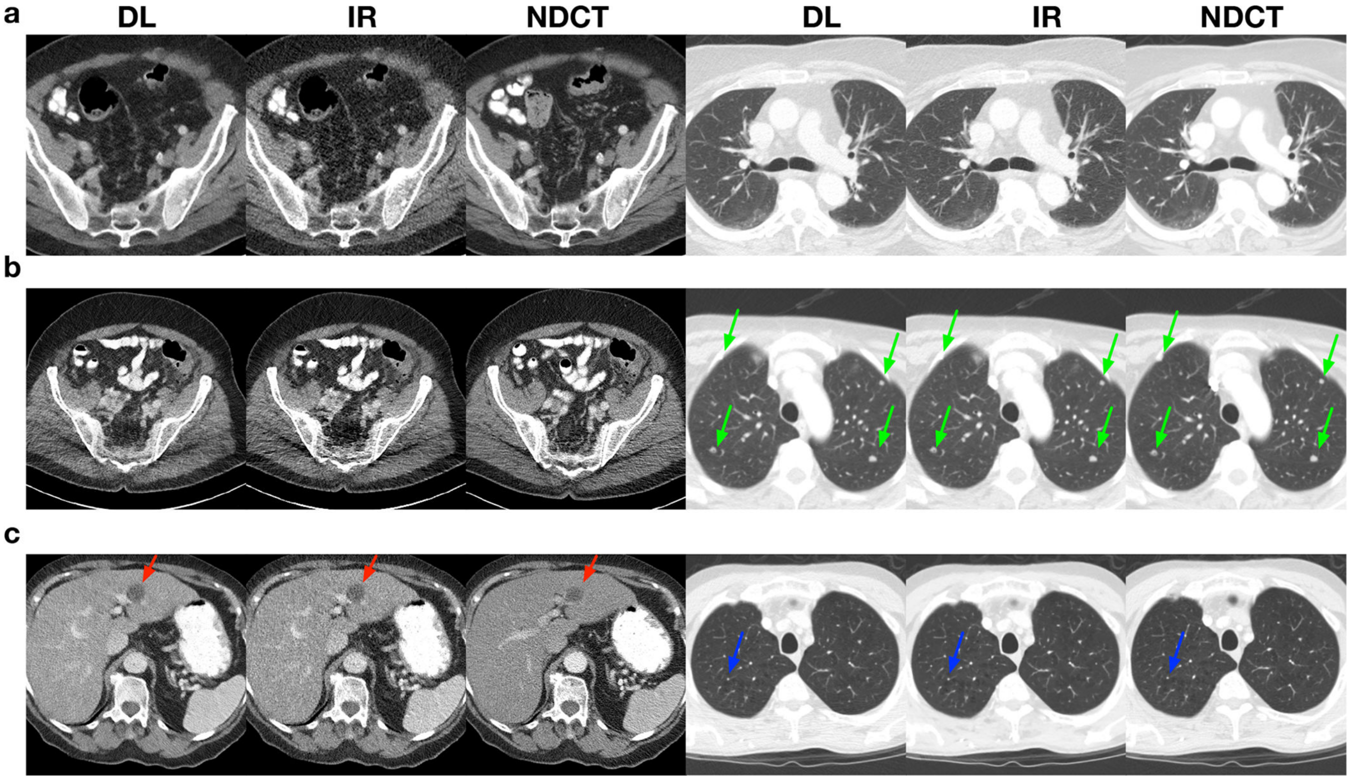

Figure 4: Best DL and best IR reconstructions as well as NDCT FBP for abdomen and chest regions from three vendors respectively.

a–c, Sample images from vendor A, vendor B, and vendor C respectively. Red, green and blue arrows indicate the liver lesion, lung nodule, and centrilobular emphysema, respectively.