

Research Article

Deep High-Resolution Network for Low-Dose X-Ray CT Denoising

Ti Bai , Dan Nguyen , Biling Wang, Steve Jiang ^{*}

Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

ARTICLE INFO

Article History

Received 31 Jan 2021
 Accepted 15 Apr 2021

Keywords

Low-dose CT
 Deep learning
 Denoise

ABSTRACT

Low-dose computed tomography (LDCT) is clinically desirable because it reduces the radiation dose to patients. However, the quality of LDCT images is often suboptimal because of the inevitable strong quantum noise. Because of their unprecedented success in computer vision, deep learning (DL)-based techniques have been used for LDCT denoising. Despite DL models' promising ability to remove noise, researchers have observed that the resolution of DL-denoised images is compromised, which decreases their clinical value. To mitigate this problem, in this work, we developed a more effective denoiser by introducing a high-resolution network (HRNet). HRNet consists of multiple branches of subnetworks that extract multiscale features that are fused together later, which substantially enhances the quality of the generated features and improves denoising performance. Experimental results demonstrated that the introduced HRNet-based denoiser outperformed the benchmarked U-Net-based denoiser, as it provided superior image resolution preservation and comparable, if not better, noise suppression. Quantitative evaluation in terms of root-mean-squared errors (RMSEs)/structure similarity index (SSIM) showed that the HRNet-based denoiser improve these values from 113.80/0.550 (LDCT) to 55.24/0.745 (HRNet), which outperformed the 59.87/0.712 for the U-Net-based denoiser.

© 2021 The Authors. Published by Atlantis Press B.V.

This is an open access article distributed under the CC BY-NC 4.0 license (<http://creativecommons.org/licenses/by-nc/4.0/>).

1. INTRODUCTION

X-ray computed tomography (CT) is widely used in clinics to visualize the internal organs of patients for diagnostic purposes. However, the radiation dose involved in X-ray CT scans constitutes a potential health concern to the human body, since it may induce genetic, cancerous, and other diseases. Therefore, following the *As Low As Reasonably Achievable* (ALARA) principle is of the utmost importance in clinics when determining the radiation dose level of a CT scan. One of the dominant ways to reduce the dose level is to decrease the exposure level of each projection angle. However, lower exposure levels inevitably introduce stronger quantum noise into the CT image, which could make accurate diagnosis and organ contouring impossible, thereby decreasing the clinical value of these images. Thus, a denoising algorithm is essential for enhancing the image quality of low-dose CT (LDCT) images.

There has been a surge of research in this direction, which can be roughly divided into three categories: projection domain-based denoising [1,2], image domain-based denoising [3,4], and regularized iterative reconstruction [5–12]. Because it is convenient to access the CT images directly from the picture archive and communication system (PACS), image domain-based denoising algorithms are becoming increasingly popular. In this work, we devote

our efforts to this last category, as we seek to enhance the image quality of LDCT.

Inspired by the unprecedented successes of convolutional neural network (CNN)-based deep learning (DL) techniques [13] in various image processing [14–18] and recognition [19–23] tasks, the medical community has widely adopted them to denoise CT images [24–31], achieving state-of-the-art denoising performance. For instance, Chen *et al.* devised a deep CNN consisting of several repeating convolutional modules that learns a mapping from LDCT images to normal-dose CT (NDCT) images [25]. To facilitate a more efficient information flow and to compensate the potential spatial resolution loss, they further proposed a new architecture that uses the well-known residual learning scheme, called residual encoder–decoder (RED) [25], and attained better denoising performance. This improvement in performance may be attributed to the more advanced architecture design. Indeed, it is well-recognized that neural network architecture is one of the dominant factors that affects an algorithm's performance, since it directly affects the qualities of the features, which are an abstract representation of the image. Moreover, the development history of CNN-based DL techniques over the past decade can be organized in terms of the evolution of the associated architectures: the original AlexNet [19] contained several convolutional layers and fully connected layers, then the VGG [32] network substantially increased the network's depth; the inception network [33] considered multiscale structures,

^{*} Corresponding author. Email: Steve.Jiang@UTSouthwestern.edu

then the well-known ResNet [22] proposed a residual connection to facilitate the information propagation; and most recently, the feature pyramid network (FPN) [34] fused the low- and high-level features. Many other architecture variants have been developed, such as SENet [35], DenseNet [36], and ResNeXt [37]. As network architectures have evolved, the associated algorithms' performance has also improved substantially, as in the case of the top-5 error rate of the well-known ImageNet 1K dataset being reduced from 20.91% (AlexNet) to 5.47% (ResNeXt).¹

The goal of our LDCT image denoising task is to suppress the noise effectively while preserving the resolution as much as possible, since noise and resolution are two important, but combating, metrics for evaluating CT image quality. High-level features are key to effectively suppressing noise by aggregating information from a large receptive field, while low-level features are essential for preserving good resolution. Therefore, extracting high-quality low- and high-level features and effectively fusing them will be extremely important when designing CNN architectures for LDCT image denoising. Toward this technical line, U-Net [38] might be one of the most well-known architectures, which add a skip connection to facilitate the fusion of the low- and the high-level features. Despite its great success in various tasks and datasets, researchers have found that applying the U-Net architecture to the LDCT denoising task still leads to significant resolution loss, thus decreasing the clinical value of the denoised CT images. Therefore, there is still room to further enhance the image quality of the denoised CT image by employing a more advanced network architecture that can extract and fuse the low- and high-level features more effectively.

In the field of natural image processing and analysis, Sun *et al.* [39] and Wang *et al.* [40] proposed a high-resolution network (HRNet)-based representation learning method that attained state-of-the-art performance across various computer vision tasks, such as human pose estimation, image classification, object detection, and segmentation. Specifically, HRNet uses multiple branches to extract multiscale features, which are then fused together internally. This architecture design can ensure a highly efficient combination of low- and high-level features during the forward propagation process and thus produce a high-quality feature for the downstream task.

To the best of our knowledge, this architecture has not been applied to denoising tasks. We feel that HRNet is highly suitable for medical image denoising tasks where high resolution is desirable to preserve the fidelity of anatomical structures. Given this, in this work, we have introduced HRNet into the medical image processing field and verified its superior denoising performance for LDCT images. We hope this newly introduced architecture will also help other researchers in the medical imaging field to boost performance on this task.

2. METHODS AND MATERIALS

2.1. Methods

Let us first mathematically formulate our problem. Given an LDCT image $x \in \mathcal{R}^N$ whose noise component is represented as $\epsilon \in \mathcal{R}^N$,

¹These data come from the model zoo provided by the official implementations of PyTorch. See <https://pytorch.org/docs/stable/torchvision/models.html> for more details.

the task of LDCT denoising is to restore the underlying clean image $y \in \mathcal{R}^N$, where both the noisy and the clean images are vectorized and N denotes the number of pixels. Without loss of generality, we can assume the noise is additive, as follows:

$$x = y + \epsilon. \quad (1)$$

From the perspective of DL, problem (1) can be regarded as an image-to-image translation problem that can be solved by learning a CNN $\Phi_W(x)$ that is parameterized by W . Given a precollected training dataset $\{(x_i, y_i) \mid i \in 0, 1, \dots, M\}$ with a size of M , where i indexes the training samples, the parameters W can be estimated by minimizing the following mean squared error (MSE)-based cost function:

$$W = \operatorname{argmin}_W \sum_i \left| \Phi_W(x_i) - y_i \right|_2^2. \quad (2)$$

In practice, one can never access a real clean image that contains no noise, since quantum noise is inevitable. Therefore, the clean image in the training dataset is usually replaced with an NDCT image \tilde{y} that is also contaminated by mild noise. In this work, we also used this relaxed but practical experimental setting to train our network.

For completeness, we illustrate the architecture of the HRNet in Figure 1. There are four different branches that extract the features from different scales. The starting feature of each scale is calculated by applying a stride-two convolutional layer to the feature of the previous scale. All features from the different scales are processed in a stage-by-stage fashion. Each stage contains two convolutional layers, as indicated by the solid arrows in Figure 1. Before proceeding to the next stage, the features from different scales are fused together. In this work, this fusion process is based on the feature summation. Specifically, if the size of the previous feature is larger than that of the current feature (such as fusing the features from branch 1 into the features from branch 3), one or more stride-two convolutional layers are used to downsample the previous feature so that it has the same size as the current feature. For example, if the previous feature has a size of 128×128 and the current feature has a size of 32×32 , then two continuous stride-two convolutional layers are used. By contrast, if the size of the previous feature is smaller than that of the current feature, then bilinear interpolation is used to upsample the previous feature. Once we have all the features from the different scales, they are concatenated together to generate the final feature. Note that features with a smaller size are first bilinearly interpolated into the same size as the original input image. A predictor is attached to the last layer to output the denoised image.

In this work, all convolutional layers except the predictor consist of three operators: a convolution operator, an instance normalization [41] operator, and the rectified nonlinear unit (ReLU). The predictor consists of the convolution operator and the ReLU. The channel number for the features in same scale is the same. Specifically, all the convolution operators in the 1st/2nd/3rd/4th branch have the same input and output channel numbers, which are 18/36/72/144, respectively. We use stride-two convolutional layers (stacked if necessary) to conduct the downsampling operation. In details, if we downsample a feature from branch i to a feature in branch j , we will use $\operatorname{ceil}\left(\frac{j-1}{i}\right)$ stride-two convolutional layers, where the convolution operator in k^{th} layer has an input channel number of $N_i/2^k$,

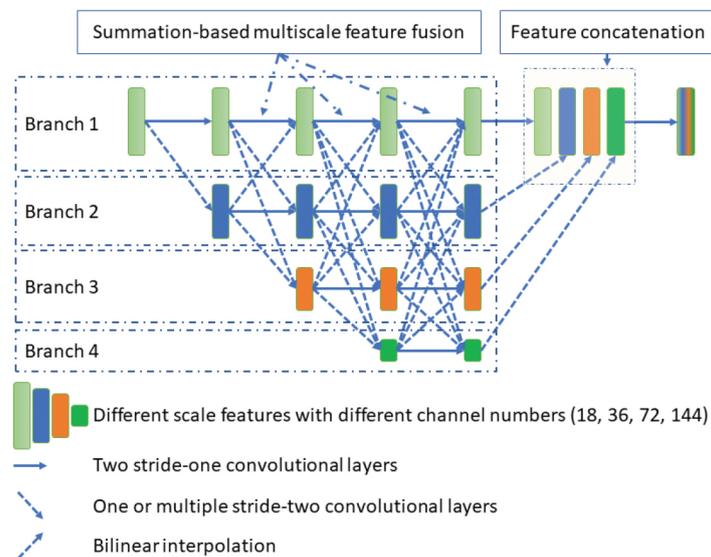


Figure 1 Illustration of the HRNet architecture. Four different branches extract the features from different scales. During the forward propagation, the features from different scales are fused together gradually. To enable the fusion process, indicated by the dashed line connections, the previous features are first downsampled or upsampled, depending on the relative size of the features, such that the summation-based feature fusion is valid. Finally, all the features from the different scales are upsampled to the original scale, if necessary, and concatenated together, resulting in the final features. A predictor is attached as the last layer to output the denoised image. All convolutional layers consist of three operators: convolution, instance normalization, and rectified nonlinear unit. The final predictor consists of two operators: convolution and rectified nonlinear unit.

and an output channel number of $N_i/2^{k+1}$. N_i denotes the channel number in the i^{th} branch. Since X-ray CT is a gray-scale image, the input and output channels are each one.

2.2. Materials

2.2.1. Training and validation datasets

In this work, we used the publicly released American Association of Physicists in Medicine (AAPM) Low-Dose CT Grand Challenge dataset [42], which consists of contrast-enhanced abdominal CT examinations, for the model training (<https://www.aapm.org/grandchallenge/lowdosect/>). Specifically, NDCT projections were first acquired according to the routine clinical protocols at Mayo Clinic, which uses the automated exposure control and automated tube potential selection strategies so that the referenced tube potential is 120 kV and the quality referenced effective exposure is 200 mAs. Then, Poisson noise was inserted into the projection data to reach a noise level that corresponded to 25% of the full dose; the resulting data were denoted as the LDCT projection data. Later, both the NDCT and the LDCT projection data were filtered back projected (FBP) into the image domain to produce the NDCT and LDCT images. The challenge host provided four different reconstructions for both dose levels with two different slice thicknesses, 1 mm and 3 mm, and two different kernels, sharp kernel or not. In this work, we used only the 1 mm sharp reconstructions to test our model's performance.

The official training dataset comprises data from ten patients. For qualitative evaluation, we randomly split these into eight and two patients to serve as the training and validation datasets, respectively. All the data involved in this work are in 2D slices. In total, there were 4800 and 1136 2D CT slice images in the training and validation datasets, respectively.

For showcase purposes, we chose four representative slices in the validation dataset to demonstrate and compare the denoising performance. The first slice was used to check the denoising performance for the abdominal site, which is the dominant site of the training dataset. The second slice corresponds to the liver, which is one of the most important human organs in the abdominal site. The third slice selected was in the lung region, which was partially covered during the CT scan. The fourth slice contains a radiologist-confirmed low-contrast lesion.

2.2.2. Testing dataset

We used patient's data from the testing dataset of the challenge to verify the model's performance. Specifically, we first rebinned the raw helical-scanned projection data into fan-beam projection data with a slice thickness of 1 mm and a pixel size of $1 \times 1 \text{ mm}^2$, which were then FBP reconstructed into the CT image domain, where we used the original ramp filter kernel with a band limited by 0.5. It should be noted that the training dataset was reconstructed with the commercial implementation of the FBP algorithm, but our testing

dataset was reconstructed with our own homemade FBP algorithm. Thus, despite having the same noise distributions in the projection domain, the image domains differ between the testing dataset and the training dataset due to different implementation details, such as different filter kernels.

2.2.3. Training details

The input and target images were the LDCT and NDCT images, respectively. In detail, we first normalized both images by dividing by 2000 so that most of the image intensities fell within the range of 0–1. Note that the pixel values of the original CT images are CT values that were shifted by 1000 HU.

To enlarge the dataset, we have conducted extensive data augmentations. Specifically, given an image of size 512×512 , we first make 16 pixels zero-padding along each image side, resulting an image with a size of 544×544 , then randomly crop out a sub-image with a size of 512×512 . Later, with a probability of 0.5, the above cropped out image is rotated around the center with a random angle ranging from -10° to $+10^\circ$.

We used the Adam [43] optimizer to minimize the cost function (2), with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The algorithm ran for 100K iterations. The learning rate was initially set as 1×10^{-4} , which was then reduced to 1×10^{-5} and 1×10^{-6} at iterations 50K and 75K, respectively. The batch size was set as 1. The PyTorch framework was used to train the network.

2.2.4. Comparison studies

For comparison, we also implemented a baseline network architecture: U-Net.

To be specific, in the encoder part, the initial output channel number in the input layer is 32. Continuous stride-two convolutional layers are applied to extract images' high-level features. The channel number is doubled per each downsampling resulting from the stride-two convolutional operators until a feature number of 512 is reached. There are nine layers in the encoder part, which leads to a bottleneck feature with a size of 1×1 , meaning that the associated receptive field covers the entire image with a size of 512×512 . In the decoder part, a concatenation operation fuses the low-level and the high-level features. The output channel number is one for the last output layer in the decoder part. As with the HRNet, the convolutional layer contains three operators: convolution, instance normalization, and ReLU.

All other training details for the U-Net were the same as for the HRNet introduced above.

2.2.5. Evaluation metrics

In this work, we evaluated the models' performance quantitatively in terms of the root-mean-square error (RMSE) and the structure similarity (SSIM) [44]. The RMSE is defined as

$$RMSE = \sqrt{|\bar{x} - \bar{y}|^2_2},$$

where \bar{x} and \bar{y} represent the evaluated image and the associated NDCT image, respectively, if available.

The SSIM is defined as

$$SSIM(\bar{x}, \bar{y}) = \frac{(2\mu_{\bar{x}}\mu_{\bar{y}} + a_1)(2\sigma_{\bar{x}}\sigma_{\bar{y}} + a_2)}{(\mu_{\bar{x}}^2 + \mu_{\bar{y}}^2 + a_2)(\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2 + a_2)},$$

where $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}^2$ represent the mean value and the variance, respectively, of the denoised image \bar{x} ; the same notations also apply for the NDCT image. We selected $a_1 = 1 \times 10^{-4}$ and $a_2 = 9 \times 10^{-4}$ to stabilize the calculation process.

We quantified the detectability of lesions via the contrast-to-noise ratio (CNR). The CNR is defined as follows:

$$CNR = \frac{2|\mu_{fg} - \mu_{bg}|}{\sigma_{fg} + \sigma_{bg}},$$

where μ_{fg} and σ_{fg} correspond to the mean value and the standard deviation, respectively, of the foreground (fg) region of interest (ROI). The same naming rules apply to the background (bg) ROI.

2.2.6. Noise analysis

As mentioned above, the model was trained with paired LDCT-NDCT images, where both the input LDCT and the target NDCT images were contaminated by noise. More specifically, the target NDCT image contained realistic quantum noise (denoted as *target noise* hereafter), while the input LDCT image contained both the target noise inherited from the NDCT image and the added simulated noise (denoted as *added noise* hereafter). We decided that it would be interesting and valuable to analyze whether the trained denoiser can remove both the target noise and the added noise.

We first defined the difference between the input LDCT and the output denoised image as *removed noise*. We used the cosine correlations between the removed noise and the added noise or the target noise to characterize the composition of the removed noise. More specifically, since the target noise is coupled with the underlying clean image, and there may be a certain number of structures in the removed noise, we calculated the cosine correlations based on the high-frequency components, where the majority is the noise. In this work, the high-frequency components are defined as those frequencies in the ranges of $[-\pi, -\frac{150}{256}\pi]$ and $[\frac{150}{256}\pi, \pi]$. Moreover, we also calculated the projection lengths from the removed noise to the target noise and the added noise to quantify how much target noise and/or added noise was removed. The projection length from vector \vec{a} to another vector \vec{b} is defined as

$$P_{\vec{b}}(\vec{a}) = |\vec{a}| \times \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \times \frac{\vec{b}}{|\vec{b}|} = \frac{\vec{a} \cdot \vec{b}}{|\vec{b}|} \frac{\vec{b}}{|\vec{b}|} \quad (3)$$

As a showcase, we used all the slices from one patient in the validation dataset to analyze the noise correlations.

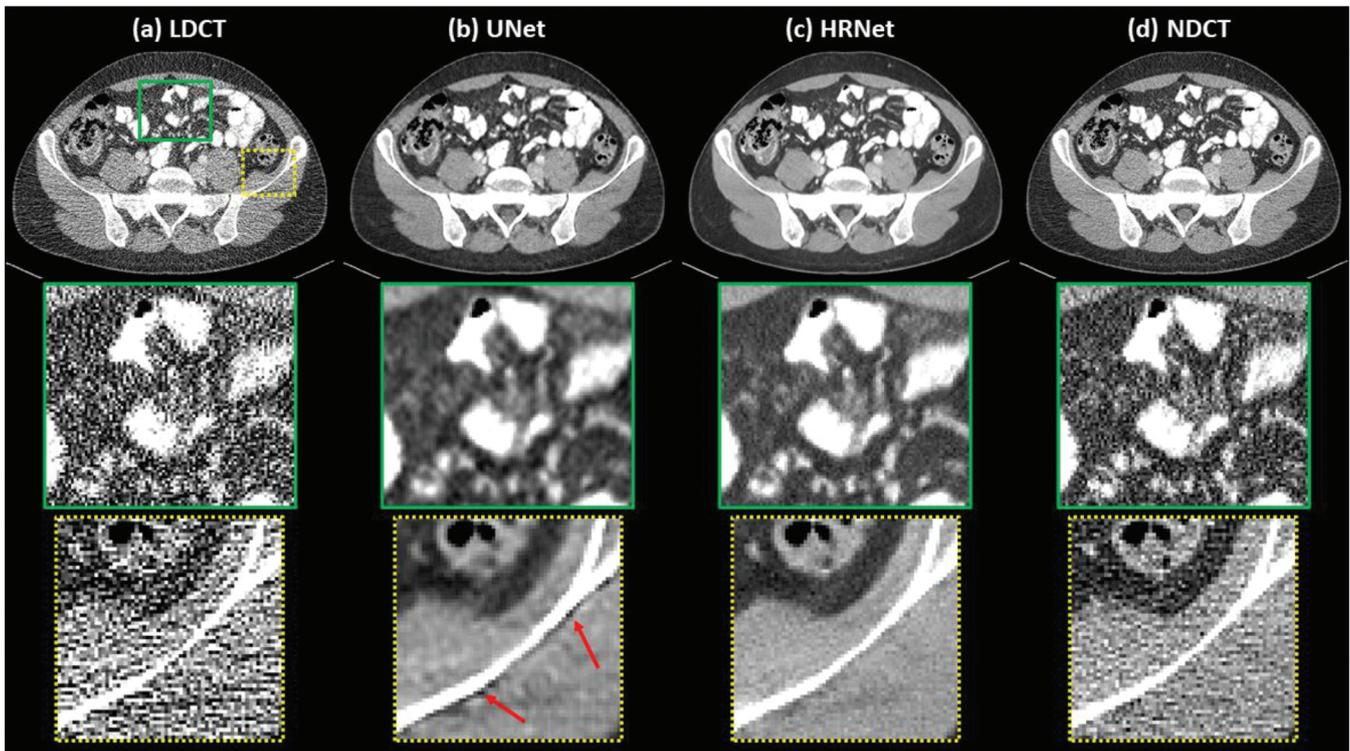


Figure 2 | Denoised results for the slice selected from the abdominal site. (a) LDCT, (b) U-Net, (c) HRNet, and (d) NDCT. Images in the second and third rows are the zoomed-in views for the contents in the first row, corresponding to the regions of interest indicated by the solid green box and the dotted yellow box, respectively. Display window: $[-160, 240]$ HU.

3. RESULTS

Figure 2 presents the denoised results for a slice in the validation dataset corresponding to the abdominal site. It is obvious that both denoisers effectively suppressed the noise, as shown in Figure 2(b) and 2(c). Compared to the NDCT image in Figure 2(d), the U-Net-based denoiser produced an image with much inferior resolution. The HRNet-based denoiser introduced here produced an image (Figure 2[c]) with much better resolution than the U-Net-based denoiser, but with slightly inferior resolution to the NDCT image. These phenomena can be clearly observed from the zoomed-in views displayed in the second row of Figure 2. In addition, the noise level of the image associated with the HRNet-based denoiser is much lower than that of the NDCT image. Another interesting finding is that using the U-Net-based denoiser led to strong artifacts around the boundary of the bone, as indicated by the red arrows in the third row of Figure 2. By contrast, the HRNet-based denoiser faithfully restored the underlying structures around that region.

Figure 3 demonstrates the denoised results for the liver organ. The strong noise in the LDCT image overwhelms the fine vessels in the liver. Both denoisers suppressed the noise effectively. Observing the zoomed-in view in the second row of Figure 3 reveals that the HRNet-based denoiser produced an image with more details, thus suggesting higher resolution, than the U-Net-based denoiser. Moreover, the noise in the image associated with the HRNet-based denoiser is weaker than in the image from the U-Net-based denoiser, which indicates the HRNet's stronger denoising ability.

The red arrows indicate that some of the vessels in the image from the HRNet-based denoiser can be more easily distinguished than even those in the NDCT image, which also suffers from the quantum noise.

Figure 4 depicts the denoised results for the lung slice. In this case, both the LDCT and the NDCT images clearly show the locations and shapes of the lung nodules. Processing the LDCT with the U-Net-based denoiser substantially decreased the resolution even though the noise was effectively removed. The image generated by the HRNet-based denoiser exhibits much weaker noise than, but slightly inferior, if not comparable, resolution to the LDCT image. Indeed, as the red arrows in Figure 4 indicate, these small lung nodules have clearer structures than in the LDCT image, which is contaminated by strong quantum noise.

Figure 5 compares different denoising results by using the slice that contains a low-contrast lesion, indicated by the solid green box. Not surprisingly, it is hard to distinguish this lesion from the surrounding normal tissues on the noisy LDCT image. Using the denoisers greatly alleviates this challenge, as shown by the images in Figure 5(b) and 5(c). Further inspection of both denoised images reveals that the HRNet-based denoiser yielded an image with a much more natural noise texture. This can be clearly seen in the zoomed-in views in the second row of Figure 5. For quantitative comparison, we also calculated the CNRs of this low-contrast lesion on different images, with results of 0.32 (LDCT), 2.45 (U-Net), 2.70 (HRNet), and 0.90 (NDCT); this further verifies the superior denoising performance of the introduced HRNet-based denoiser.

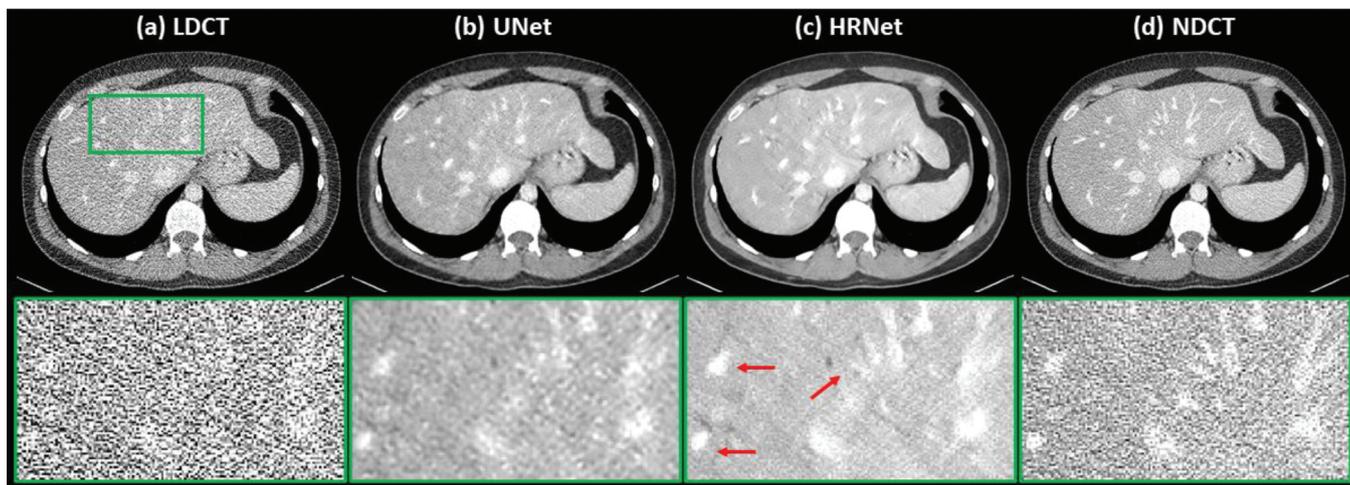


Figure 3 | Denoised results for the slice corresponding to the liver organ. (a) LDCT, (b) U-Net, (c) HRNet, and (d) NDCT. Images in the second row are the zoomed-in views for the contents in the first row, corresponding to the region of interest indicated by the solid green box. Display window: $[-160, 240]$ HU.

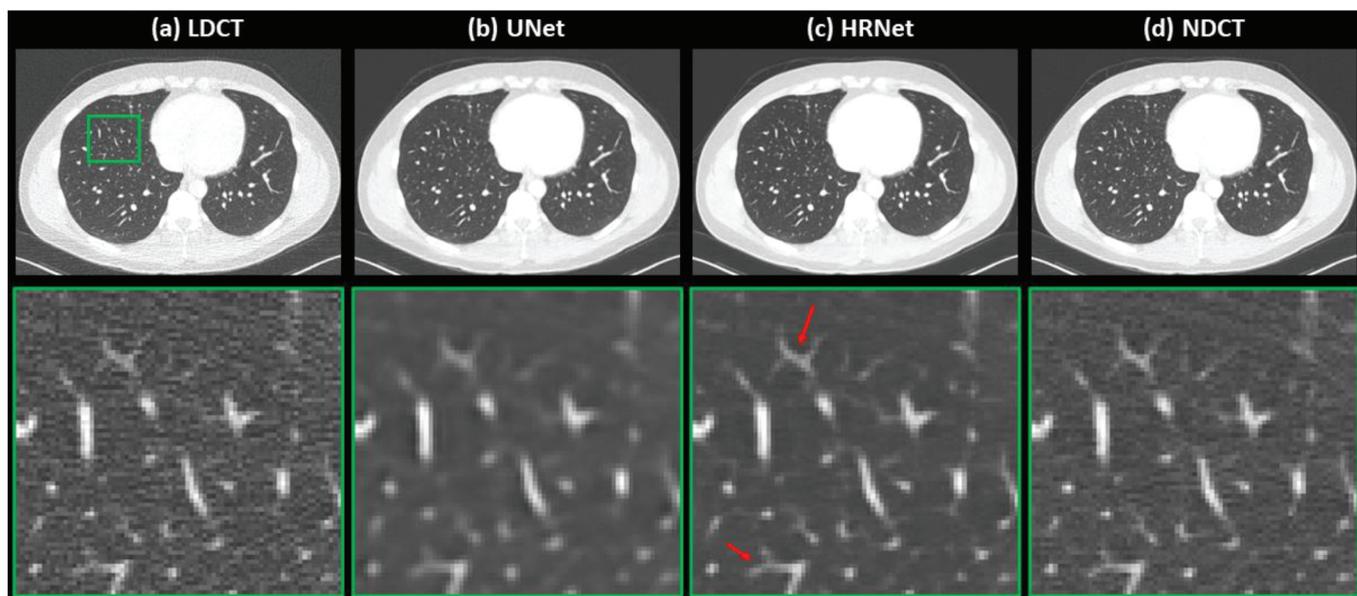


Figure 4 | Denoised results for the slice corresponding to the lung region. (a) LDCT, (b) U-Net, (c) HRNet, and (d) NDCT. Images in the second row are the zoomed-in views for the contents in the first row, corresponding to the region of interest indicated by the solid green box. Display window: $[-160, 240]$ HU.

To further quantify the performance of both denoisers, we calculated the RMSE and SSIM values for each showcase in Figures 2–5 with their reference NDCT images, as listed in Table 1. We can see that both denoisers dramatically reduced the RMSE values, but the HRNet-based denoiser yielded a greater improvement than the U-Net-based denoiser. Similar findings can also be observed from the SSIM-based metrics. For a more comprehensive comparison, we calculated and plotted the RMSE and SSIM values for all the slices in the validation dataset (Figure 6). This plot shows that the introduced HRNet-based denoiser outperformed the U-Net-based denoiser in almost every evaluated slice, even though both denoisers substantially enhanced the image quality. For an overall quantitative comparison, we also computed the averaged RMSE and SSIM values for the entire validation dataset. As tabulated in Table 1,

the RMSE decreased from 113.80 HU (LDCT) to 59.87 HU when applying the U-Net-based denoiser, and it decreased further to 55.24 HU with the HRNet-based denoiser. Again, the SSIM values also validate the image quality enhancement, as they increased from 0.550 (LDCT) to 0.712 (U-Net) and 0.745 (HRNet).

To further investigate the denoising performance boundary of the introduced HRNet-based denoiser, in Figure 7, we demonstrate two worst cases in terms of the lowest SSIM values for the two patient cases in the validation dataset. It clearly reveals that the proposed HRNet-based denoiser still outperforms the U-Net-based denoiser, with better noise suppression and superior details preservation ability. This can be also verified from the quantitative metrics as tabulated in Table 2.

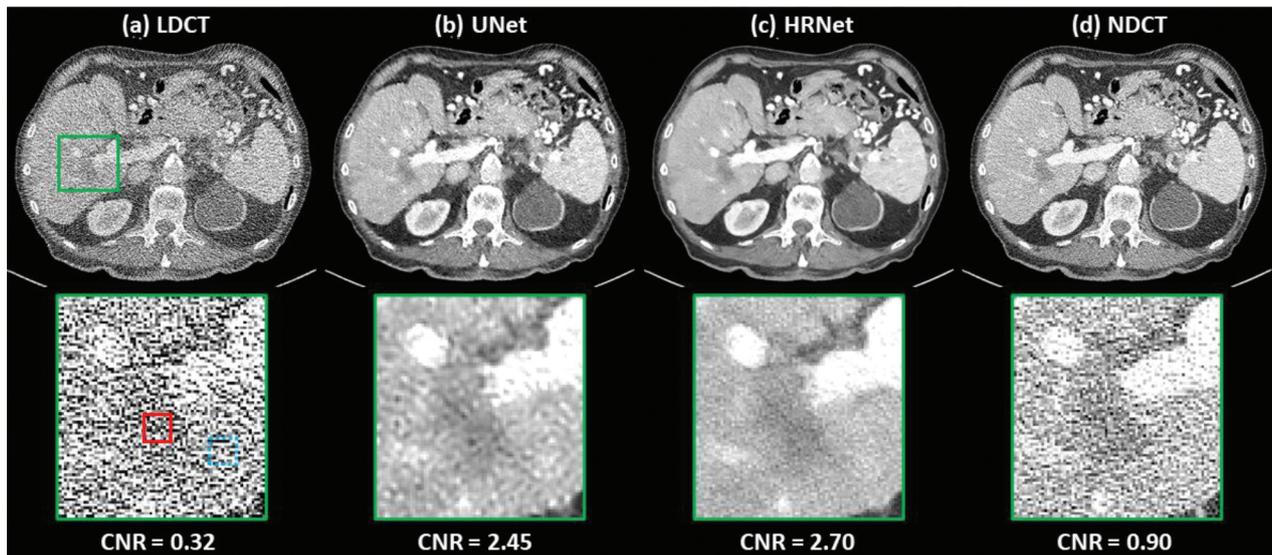


Figure 5 | Denoised results for the slice containing a low-contrast lesion, as indicated by the solid green box. (a) LDCT, (b) U-Net, (c) HRNet, and (d) NDCT. Images in the second row are the zoomed-in views of the low-contrast lesion. The red solid box and the blue dotted box indicate the foreground and background contents, respectively, used for the CNR calculation, and the associated results are shown below. Display window: $[-160, 240]$ HU.

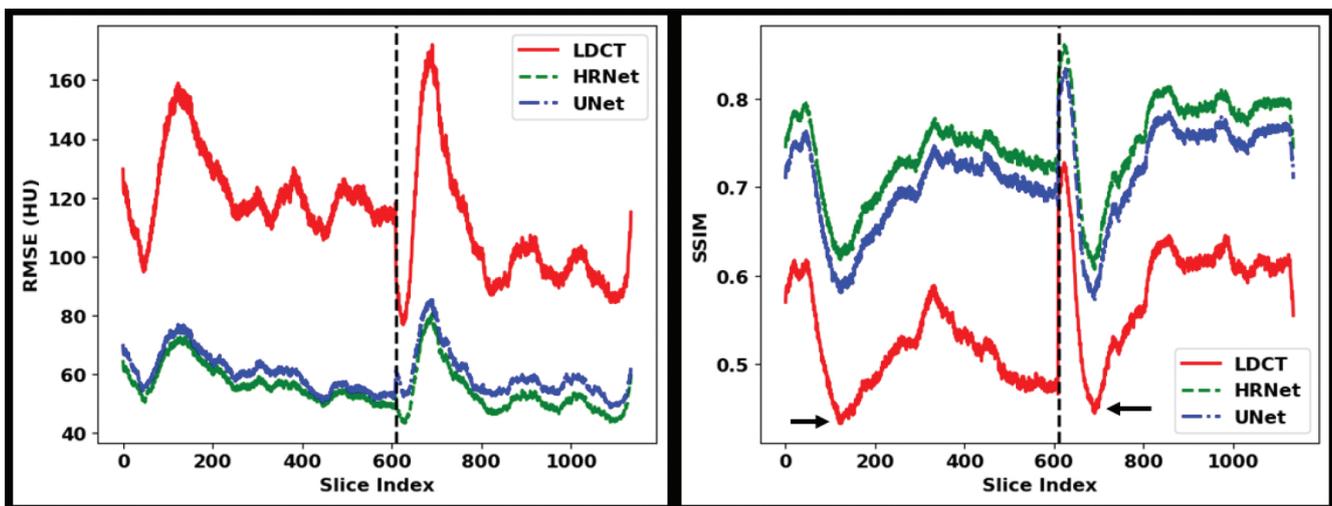


Figure 6 | Quantitative results for each slice in the validation dataset. The x-axes are the slice index and the y-axes are the RMSE (left figure) and the SSIM (right figure). The solid red, dashed green, and dotted dash blue curves pertain to the LDCT, the HRNet method, and the U-Net method, respectively. The slices of the two patients in the validation dataset are separated by the dashed black vertical line.

Table 1 | Quantitative results in terms of RMSE and SSIM for the four showcases in Figures 2–5, as well as the whole validation dataset. The RMSE has a unit of HU.

		Figure 2	Figure 3	Figure 4	Figure 5	Dataset
RMSE (HU)	LDCT	115.49	124.67	120.90	164.52	113.80
	U-Net	61.08	63.84	66.65	82.32	59.87
	HRNet	56.10	60.55	61.01	76.94	55.24
SSIM	LDCT	0.571	0.516	0.588	0.451	0.550
	U-Net	0.734	0.666	0.727	0.584	0.712
	HRNet	0.766	0.702	0.761	0.620	0.745

The denoising results for a slice in the testing dataset are shown in Figure 8. We can see that the HRNet-based denoiser delivered an

image with sharper edges and a comparable, if not weaker, noise level than the U-Net-based denoiser.

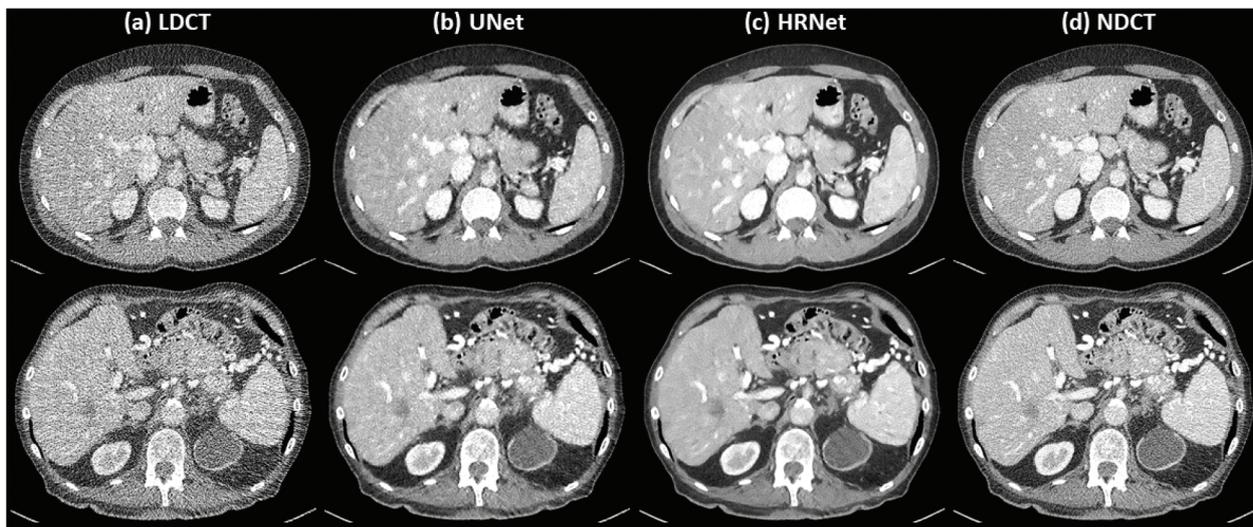


Figure 7 | Denoised results for the two slices that exhibits the lowest SSIM values in the two validation patients. (a) LDCT, (b) U-Net, (c) HRNet, and (d) NDCT. The images in the first and the second rows correspond to the slices indicated by the left and the right arrows in Figure 6 (SSIM subfigure). Display window: [-160, 240] HU.

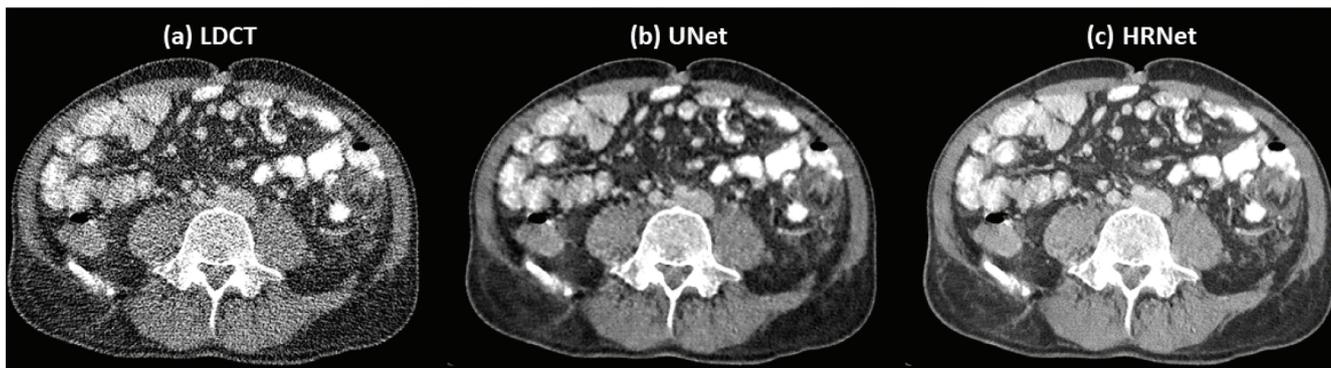


Figure 8 | Denoised results for the slice in the testing dataset. (a) LDCT, (b) U-Net, and (c) HRNet. Display window: [-160, 240] HU.

Table 2 | Quantitative results in terms of RMSE and SSIM for the two worst showcases in Figure 7 from the two patient cases in the validation dataset. The RMSE has a unit of HU.

Figure 7		Top Row	Bottom Row
RMSE (HU)	LDCT	153.61	164.67
	U-Net	74.49	82.87
	HRNet	70.71	77.60
SSIM	LDCT	0.441	0.454
	U-Net	0.692	0.584
	HRNet	0.629	0.620

The noise analysis is shown in Figure 9. We can observe that the difference image (Figure 9[b]) between the input and the denoised images contains almost exclusively noise, which suggests that the introduced HRNet-based denoiser preserved structural details very well. The corresponding Fourier domain depicted in Figure 9(c) suggests that the noise is dominated by both the low- and the high-frequency components, while the middle-frequency noise has been effectively removed. After we masked out the low-frequency part (Figure 9[f]) and inversely transformed back to the image domain

(Figure 9[e]), the noise contained only the high-frequency component, which can be clearly observed by comparing Figure 9(b) and 9(e). Since the low-frequency component is excluded, the noise power in Figure 9(e) is lower than in Figure 9(b). The cosine correlations among the different noise components are plotted in Figure 9(d). We can see that the target noise and the extra added noise are almost orthogonal, with a cosine correlation of around -0.08, which is close to zero. The correlation is around 0.9 between the removed noise and the extra added noise and around 0.3 between the removed noise and the target noise. These two values indicate that most of the removed noise comes from the extra added noise, though the denoiser also removes some noise from the target image. The projection length from the removed noise to the extra added noise is 78.46%, and the projection length from the removed noise to the target noise is 61.31%. After we subtract the projected noise from the removed noise, the remaining noise has an energy of 11.37% of all the removed noise.

4. DISCUSSIONS AND CONCLUSIONS

The goal of any CT denoiser is to suppress the noise as much as possible while preserving anatomical details as well as possible.

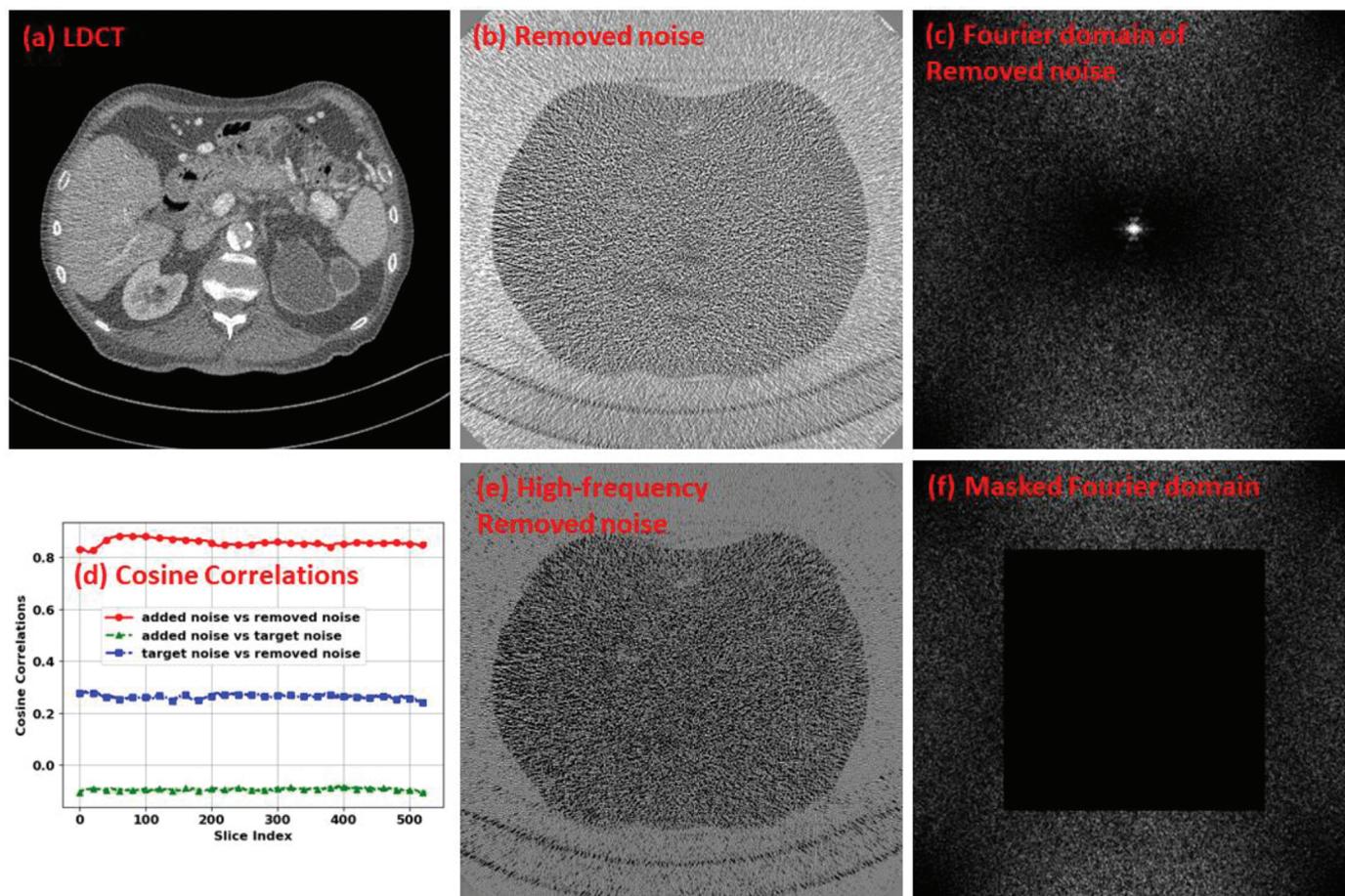


Figure 9 Noise analysis. (a) LDCT, (b) noise removed by the HRNet-based denoiser, (c) Fourier domain of the removed noise, (d) cosine correlations among the added noise, the removed noise and the target noise, (e) the high-frequency components of the removed noise, and (f) the masked Fourier domain. The display window is $[-160, 240]$ HU for image (a), $[-50, 50]$ HU for images (b) and (e), and $[10^4, 10^5]$ HU for images (c) and (f).

DL-based denoisers have proven very effective at suppressing noise by automatically extracting image features, thus achieving state-of-the-art denoising performance. However, the feature quality highly depends on the model architecture. For the denoising task, both the low- and the high-level features are important. The former is important for preserving details, while the latter is essential for effectively suppressing noise by using context information on a large scale. The encoder–decoder architecture is efficient for high-level feature extraction, but it lacks low-level information. The U-Net can improve the low-level feature quality to some extent by using skip connections, but still cannot provide a high low-level feature quality that can be used to faithfully restore the fine details. This deficiency can be clearly observed from Figure 2 where there exist strong dark artifacts around the boundary of the bone regarding the U-Net-based denoising result, as indicated by the red arrows. This extracted low-quality low-level feature probably also explain why the U-Net-based denoiser leads to oversmoothed structures despite it can suppress the noise effectively. By contrast, the introduced HRNet generated both low- and high-level features of high quality by using different branches to extract the features from different levels, then fusing the different level features together. Consequently, our experimental results verified the superiority of this architecture design and showed that HRNet can effectively remove noise while preserving the fine anatomical structures very well.

For some cases, the results from the HRNet were even better than the NDCT, such as for the low-contrast lesion detection task (Figure 5). This might be because we removed not only the added simulated noise but also the noise inherited from the target NDCT image. The noise analysis above supports our hypothesis, as it shows that the HRNet-based denoiser removed 61.31% of the target noise from the NDCT image. Nonetheless, we are not claiming that the HRNet-based denoiser can deliver an image whose quality surpasses that of NDCT. Actually, when visualizing high-contrast details that are robust to the noise but sensitive to the resolution, such as the lung nodules shown in Figure 4, the structures associated with the HRNet are slightly oversmoothed compared to the NDCT.

It is well known that data-driven DL models may suffer from the problem of model generalizability when there is a distribution gap between the training and the testing environments. In this work, the models were tested only on simulated datasets whose data distributions are similar to the training dataset, if one ignores the potential differences caused by different reconstruction parameters. More realistic testing datasets are required to evaluate the model's performance further before it can be translated to the clinic.

In summary, in this work, we introduced an HRNet-based denoiser to enhance the quality of LDCT images. Because it extracts both

low- and high-level features of high quality, the HRNet can deliver an enhanced image with effectively suppressed noise and well preserved details. Compared to the U-Net-based denoiser, the HRNet can produce images with higher resolution. Quantitative experiments showed that the introduced HRNet-based denoiser can improve the RMSE/SSIM values from 113.80/59.87 (LDCT) to 55.24/0.745, and it outperformed the U-Net-based denoiser, whose values are 59.87/0.712.

DATA AVAILABILITY

The data used in this paper can be publicly accessed from the official website of the AAPM Low-Dose CT Challenge (<https://www.aapm.org/grandchallenge/lowdosect/>).

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Steve Jiang: Initiated the project; Ti Bai, Biling Wang, Dan Nguyen and Steve Jiang: Designed the experiments; Ti Bai: Performed the model training; Biling Wang: Conducted the data collection and the data analysis; Ti Bai: Wrote the paper; Steve Jiang: Edited the paper.

ACKNOWLEDGMENTS

We would like to thank Varian Medical Systems Inc. for supporting this study and Dr. Jonathan Feinberg for editing the manuscript. We also would like to thank Dr. Ge Wang from Rensselaer Polytechnic Institute for his constructive discussions and comments.

REFERENCES

- [1] J. Wang, *et al.*, Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters, in *Medical Imaging 2005: Image Processing*, International Society for Optics and Photonics, San Diego, CA, USA, 2005.
- [2] A. Manduca, *et al.*, Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT, *Med. Phys.* 36 (2009), 4911–4919.
- [3] K. Dabov, *et al.*, Image denoising with block-matching and 3D filtering, in *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, International Society for Optics and Photonics, San Jose, CA, USA, 2006.
- [4] K. Dabov, *et al.*, Image denoising by sparse 3-D transform-domain collaborative filtering, *IEEE Trans. Image Process.* 16 (2007), 2080–2095.
- [5] T. Bai, *et al.*, Z-index parameterization for volumetric CT image reconstruction via 3-D dictionary learning, *IEEE Trans. Med. Imaging.* 36 (2017), 2466–2478.
- [6] H. Yan, *et al.*, Towards the clinical implementation of iterative low-dose cone-beam CT reconstruction in image-guided radiation therapy: cone/ring artifact correction and multiple GPU implementation, *Med. Phys.* 41 (2014), 111912.
- [7] I.A. Elbakri, J.A. Fessler, Statistical image reconstruction for polyenergetic X-ray computed tomography, *IEEE Trans. Med. Imaging.* 21 (2002), 89–99.
- [8] J. Wang, *et al.*, Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography, *IEEE Trans. Med. Imaging.* 25 (2006), 1272–1283.
- [9] G.H. Chen, J. Tang, S. Leng, Prior Image Constrained Compressed Sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets, *Med. Phys.* 35 (2008), 660–663.
- [10] E.Y. Sidky, X. Pan, Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization, *Phys. Med. Biol.* 53 (2008), 4777.
- [11] J. Wang, T. Li, L. Xing, Iterative image reconstruction for CBCT using edge-preserving prior, *Med. Phys.* 36 (2009), 252–260.
- [12] Q. Xu, *et al.*, Low-dose X-ray CT reconstruction via dictionary learning, *IEEE Trans. Med. Imaging.* 31 (2012), 1682–1697.
- [13] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature.* 521 (2015), 436–444.
- [14] J. Caballero, *et al.*, Real-time video super-resolution with spatio-temporal networks and motion compensation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2016, pp. 2848–2857.
- [15] A. Kappeler, *et al.*, Video super-resolution with convolutional neural networks, *IEEE Trans. Comput. Imaging.* 2 (2016), 109–122.
- [16] X.J. Mao, C. Shen, Y.B. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, 2016.
- [17] D. Ulyanov, A. Vedaldi, V. Lempitsky, Deep Image Prior, In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 9446–9454.
- [18] J. Lehtinen, *et al.*, Noise2noise: learning image restoration without clean data, In *ICML*, Stockholm, Sweden, 2018. arXiv preprint arXiv:1803.04189, 2018.
- [19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, 2012, pp. 1097–1105.
- [20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014). arXiv preprint arXiv:1409.1556.
- [21] R. Girshick, Fast R-CNN, in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.
- [22] K. He, *et al.*, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2015.
- [23] K. He, *et al.*, Mask R-CNN, *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2020), 386–397.
- [24] H. Chen, *et al.*, Low-dose CT with a residual encoder-decoder convolutional neural network, *IEEE Trans. Med. Imaging.* 36 (2017), 2524–2535.
- [25] H. Chen, *et al.*, Low-dose CT via convolutional neural network, *Biomed. Optics Express.* 8 (2017), 679–694.
- [26] H. Shan, *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 (2018), 1522–1534.

- [27] Q. Yang, *et al.*, Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss, *IEEE Trans. Med. Imaging.* 37 (2018), 1348–1357.
- [28] C. You, *et al.*, Structurally-sensitive multi-scale deep neural network for low-dose CT denoising, *IEEE Access.* 6 (2018), 41839–41855.
- [29] H. Shan, *et al.*, Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction, *Nat. Mach. Intell.* 1 (2019), 269–269.
- [30] T. Bai, *et al.*, Probabilistic self-learning framework for low-dose CT denoising, [arXiv:2006.00327](https://arxiv.org/abs/2006.00327), 2020.
- [31] G. Wang, S. Li, Low-dose CT image denoising using parallel-clone networks, [arXiv:2005.06724v1](https://arxiv.org/abs/2005.06724v1), 2020.
- [32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, [arXiv:1409.1556](https://arxiv.org/abs/1409.1556), 2014.
- [33] C. Szegedy, *et al.*, Inception-v4, inception-ResNet and the impact of residual connections on learning, In *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017.
- [34] T.-Y. Lin, *et al.*, Feature pyramid networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016.
- [35] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.
- [36] G. Huang, *et al.*, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2016.
- [37] S. Xie, *et al.*, Aggregated residual transformations for deep neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.
- [38] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015.
- [39] K. Sun, *et al.*, Deep high-resolution representation learning for human pose estimation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.
- [40] J. Wang, *et al.*, Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).
- [41] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: the missing ingredient for fast stylization, [arXiv preprint arXiv:1607.08022](https://arxiv.org/abs/1607.08022), 2016.
- [42] B. Chen, *et al.*, Development and validation of an open data format for CT projection data, *Med. Phys.* 42 (2015), 6964–6972.
- [43] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, (2014). [arXiv preprint arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [44] W. Zhou, *et al.*, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (2004), 600–612.