RIDnet: Radiologist-Inspired Deep Neural Network for Low-dose CT Denoising

Kecheng Chen, *Student Member, IEEE*, Jiayu Sun, Jiang Shen, Jixiang Luo, Xinyu Zhang, Xuelin Pan, Dongsheng Wu, Yue Zhao, Miguel Bento, Yazhou Ren and Xiaorong Pu

***Abstract*—Being low-level radiation exposure and less harmful to health, low-dose computed tomography (LDCT) has been widely adopted in the early screening of lung cancer and COVID-19. LDCT images inevitably suffer from the degradation problem caused by complex noises . It was reported that, compared with commercial iterative reconstruction methods, deep learning (DL)-based LDCT denoising methods using convolutional neural network (CNN) achieved competitive performance. Most existing DL-based methods focus on the local information extracted by CNN, while ignoring both explicit non-local and context information (which are leveraged by radiologists). To address this issue, we propose a novel deep learning model named radiologist-inspired deep denoising network (RIDnet) to imitate the workflow of a radiologist reading LDCT images. Concretely, the proposed model explicitly integrates all the local, non-local and context information rather than local information only. Our radiologist-inspired model is potentially favoured by radiologists as a familiar workflow. A double-blind reader study on a public clinical dataset shows that, compared with state-of-the-art methods, our proposed model achieves the most impressive performance in terms of the structural fidelity, the noise suppression and the overall score. As a physicians-inspired model, RIDnet gives a new research roadmap that takes into account the behavior of physicians when designing decision support tools for assisting clinical diagnosis. Models and code are available at https://github.com/tonyckc/RIDnet_demo.**

***Index Terms*—LDCT denoising, deep learning, radiologist-inspired, graph convolution**

I. Introduction

COMPUTED tomography (CT) is one of the most frequently used imaging technologies in modern medicine [1]. Compared with conventional radiography, CT has the advantages of superior contrast resolution [2], superb detailed anatomical representations [3] and the ability to selectively enhance or remove structures from images [4]. However, recent studies report that the radiation exposure of CT scans may come with potential cancer risks, especially for children [5]. In the past two decades, low-dose computed tomography (LDCT) thus has become a hot screening tool for noninvasive low radiation examination, such as the early detection of lung cancer [5] and the diagnosis of COVID-19 pneumonia [6]. The reduction of dose resulting in heavily noisy CT images is a though challenge, since it will affect the diagnostic accuracy for the radiologists.

To overcome this problem, quite a few studies make an attempt to obtain the latent noise-free CT images by removing noise from LDCT images [7]–[10]. Existing LDCT denoising methods can be roughly divided into three streams. The first two streams are *sinogram filtration* [11]–[13] and *iterative reconstruction* [8], [14] based methods, respectively. Both of them achieve an effective denoising performance on account of involving the projection data directly. It is also an intractable problem that the acquisition of projection data is fairly complicated in clinical environments [9]. In addition, the *iterative reconstruction* based methods need to transform the data from the projection domain to image domain constantly [15], which is usually regarded as a time consuming process [16]. Thanks to the progress of deep neural network, the *post-processing* based methods using convolutional neural networks (CNN) [7], [10], [15], [17] realize the best denoising performance compared with the first two streams. Instead of relying heavily on projection data, such deep learning-based LDCT denoising methods work in the image domain of CT data directly [18], which is extremely convenient. The time consumption of deep learning-based methods is significantly lower than the first two streams [7].

For existing deep learning-based LDCT denoising methods using CNN, the denoising task is usually regarded as a mapping [18], [19] or texture transfer [9], [15], [20] from LDCT images to corresponding NDCT images. They often utilize various CNN-based models. However, litter attention has been paid to the inner mechanism of adopted models, such as what information has been learned by the model for LDCT images denoising? To the best of our knowledge, CNN-based models concentrate more on the extraction of local information [21]. This denoising mechanism differs from that of the one used by radiologists, resulting in a potential problem in clinical use. Radiologists are more inclined to trust a model inspired by their own behavior or workflow when reading the LDCT images, rather than a significantly different model compared with their work mechanism. Interestingly, what is a radiologist's reading behavior or workflow?

Through an in-depth communication with radiologists, the LDCT denoising mechanism of radiologists is typically a three-step workflow, which can be roughly summarized as shown in Figure 1. First, the radiologists will focus on the

This work was supported in part by Sichuan Science and Technology Program (Nos. 2020YFS0119 and 2021YFS01 72) and National Natural Science Foundation of China (No. 61806043).

K. Chen, J. Luo, X. Zhang, Y. Zhao, M. Bento, Y. Ren, and X. Pu are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China.

J. Sun and X. Pan are with the Radiology Department, West China Hospital, Sichuan University, Chengdu, 610041, China.

J. Shen and D. Wu are with the Radiology Department, West China No.4 Hospital, Sichuan University, Chengdu, 610041, China.

Co-corresponding authors: Xiaorong Pu and Yazhou Ren (e-mails: puxiaor@uestc.edu.cn; yazhou.ren@uestc.edu.cn;). K. Chen and J. Sun contributed equally to this paper.

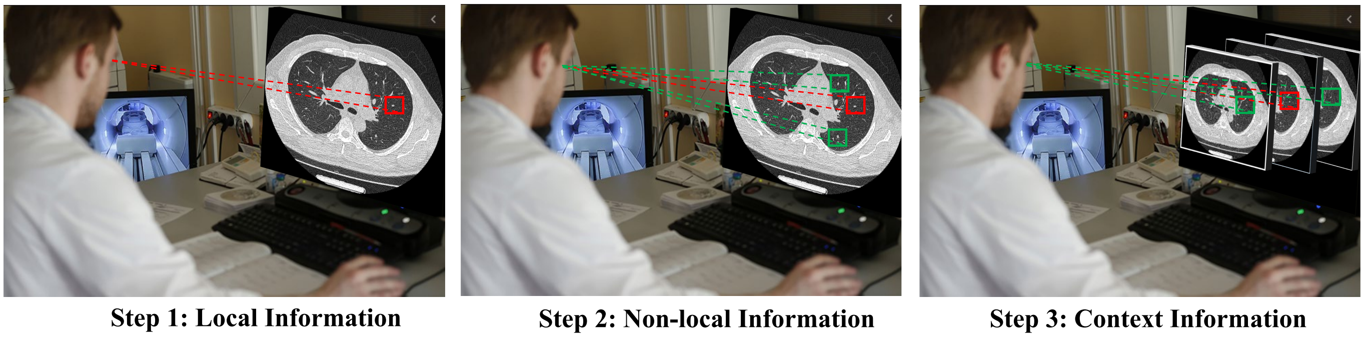

Step 1: Local Information
Step 2: Non-local Information
Step 3: Context Information

Fig. 1. The workflow of radiologists when reading the LDCT images. In step 1, the radiologist focuses on the region of interest (ROI) with local information. In step 2, the radiologist leverages those non-local but easily-observed similar tissues to perform auxiliary observation. In step 3, the radiologist slides the mouse so that the similar ROI in the front and rear slices can be observed.

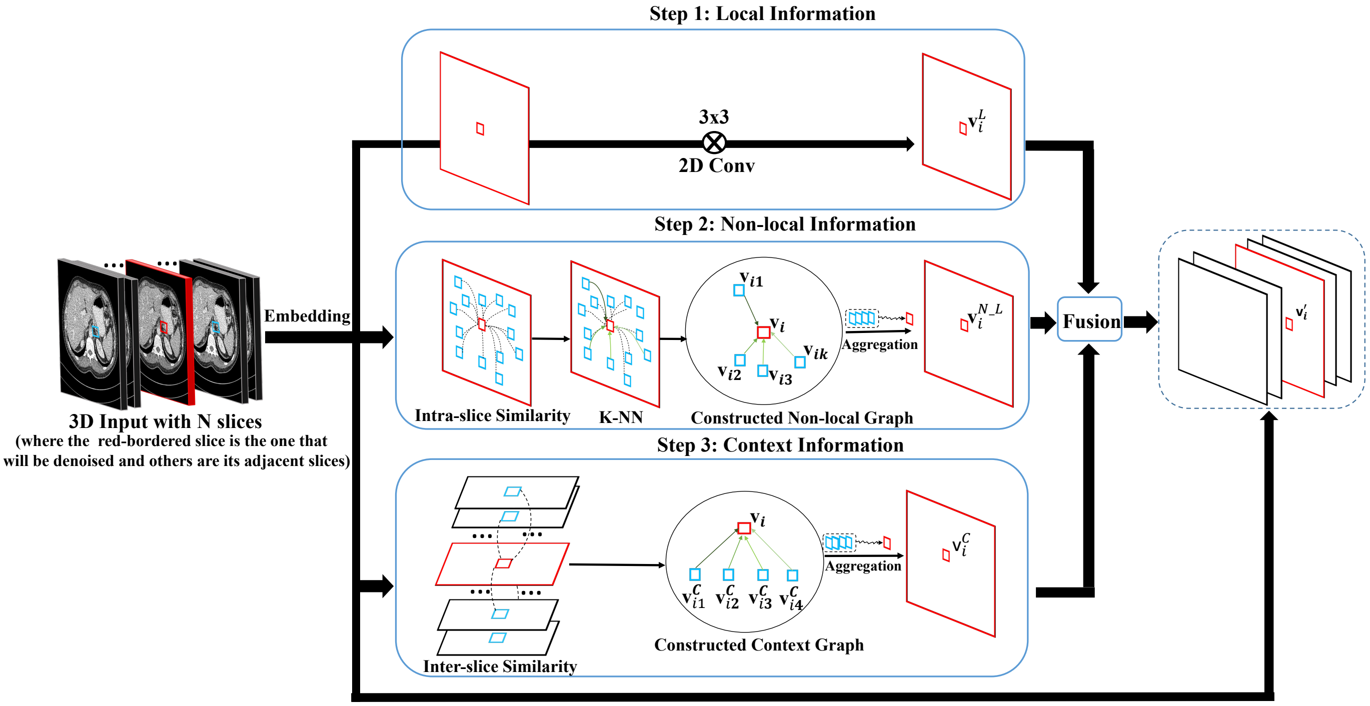


Fig. 2. The proposed radiologist-inspired deep denoising network (RIDnet). Followed by extraction of the feature map, the classical 2D convolution operation is applied in step 1 and the computation of the intra-slice similarity, the computation of a K nearest neighbor (K-NN), the construction of the non-local graph, as well as the feature aggregation is successively performed in step 2. In step 3, the whole feature maps are inputted in and then compute the inter-slice, the construction of the context graph and the feature aggregation. Finally, the result of feature fusion is used to replace the feature map in original input position, in order to keep the original 3D shape [16]. Please see the section Method for more details.

region of interest (ROI) which represents the local information (illustrated as the red box in the step 1 of Figure 1), such as tiny bronchus of the lung and subtle nodules of the lung. However, such subtle tissue structure and blood vessel will be converted by the noise easily, making it hard to be observed for radiologists. Second, to further read these hardly-observed tissues, the radiologists will leverage those non-local but easily-observed similar tissues to perform auxiliary observation (illustrated as the green boxes in the step 2 of Figure 1). This step shows the importance of non-local information. Third, it is the natural behavior for radiologists to slide the mouse so that the similar ROI in the front and rear slices can be observed (illustrated as the green boxes in the step 3 of Figure 1). The heavily noisy ROI may have a noise-free counterpart in the front or rear slice, because of the randomness of noise in different slices. This implicitly leverages the context information. According to aforementioned discussions, we can conclude that the reading workflow of radiologists adopts comprehensive information, i.e., local information, non-local information and context information, to obtain the best observation from the noisy ROI.

Conversely, existing deep learning-based LDCT denoising methods rely heavily on the convolutional neural network (CNN) [7], [17], [22]. The convolution, essentially, is a local operator that extracts some useful local information within it's filter size. Although these CNN-based methods achieve impressive performance, the explicit non-local and context information (which are also leveraged by the radiologist as mentioned in aforementioned discussions) are typically ignored by most existing methods. The most important issue that should be noted is, radiologists may not support a model whose workflow is vastly different from theirs.

To this end, we propose a novel deep learning model, namely radiologist-inspired deep denoising network (RIDnet), to imitate the workflow of a radiologist reading LDCT images. As shown in Figure 2, following the scheme of radiologist-inspired comprehensive information, we first introduce the graph convolutional network (GCN) to imitate the step 2 of radiologists through constructing the non-local graph, i.e., explicitly extracting non-local information, which cannot be utilized by existing CNN-based methods directly. Furthermore, the GCN is also proposed to imitate the step 3 of radiologists via establishing the context graph among slices for context information. Compared with the context information implicitly extracted by 3D CNN [17], [21], the information extracted by GCN is more explicit, leading to better effects as suggested in [23]. Here, the imitation of step 2 and step 3 together is composed of a novel deep learning module, namely 3D GCN module. The proposed 3D GCN module aims to imitate the behavior of radiologists, i.e., concerning the non-local information of the intra-slice and the context information of the inter-slice. Obviously, the local information is also important. We therefore leverage classical convolution operation to acquire useful local patterns. Finally, the proposed RIDnet integrates the aforementioned three parts to learn the best adaptive composition through the feature fusion. Note that the proposed RIDnet can be also stacked to further improve the denoising capacity. In this paper, the adopted model is constructed by three RIDnets due to the consideration of comprehensive performance (we discuss this in supplement I), as shown in Figure 3. Our proposed RIDnet follows the workflow of radiologists for LDCT images denoising, differentiating from the widely adopted CNN.

II. Materials and methods

A. PRELIMINARY

First, some preliminaries need to be claimed. Given that an LDCT-NDCT paired dataset, T=$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where N denotes the number of paired training samples. $(\mathbf{x}, \mathbf{y}) \in (\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are two image domains, respectively. The paired samples $\mathbf{x} \in \mathbb{R}^{n \times n}$ and $\mathbf{y} \in \mathbb{R}^{n \times n}$ are matrix expression of a noisy LDCT image and a high-quality NDCT image. Assume that the ith slice $\mathbf{x}_i$ is the one that needs to be denoised in overall slices of a patient. The 3D input of $\mathbf{x}_i$ can be represented as $\mathbf{X}_i = concat(\mathbf{x}_{-s}, \mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_s)$, where $2s+1$ is a predefined number of slices and set to 3 as suggested in [21]. The *concat* operator denotes the concatenation operator along with the first dimension.

B. The methodology of proposed RIDnet model

The proposed RIDnet model consists of four parts, i.e., the layer of embedding, the layer of extracting local information, the 3D graph convolutional networks (3D GCN), and the layer of feature fusion. We will discuss them in details.

1) The layer of embedding.: It is usually more effective for model learning to convert the input into it's embedding in feature space [24]. Every pixel then can be regarded as the format of the feature vector along with the channel direction. To this end, the classical 3D convolution is adopted in our proposed RIDnet model. Specifically, the first RIDnet model has two 3D convolution layers with channels of 64 and 32, due to the need of stronger capacity of feature extraction in the start of a model. Other RIDnet models only have one 3D convolution layer with channel of 32. To maintain the size of the feature map, the operation of 'reflect' padding is applied, which is useful to avoid the artifacts of the edge as suggested in [24]. The output of embedding layer thus can be denoted as $\mathbf{X'}_i = (\mathbf{x'}_{-s}, \mathbf{x'}_{i-1}, \mathbf{x'}_i, \mathbf{x'}_{i+1}, \cdots, \mathbf{x'}_s) = embedding(\mathbf{X}_i)$.

2) The 3D graph convolutional networks.: As the most important part in the proposed RIDnet model, the 3D graph convolutional networks (3D GCN) consists of two modules, i.e., plane graph convolutional module, depth graph convolutional module, respectively.

3) Plane Graph Convolutional Module.: For every pixel in the feature space of a CT slice, let construct a plane graph $G_p = (V, E)$. Note that we only extract non-local information of feature map that needs to be denoised, i.e., the $\mathbf{x'}_i$ of $\mathbf{X'}_i$, due to the considerations of efficiency and the contribution to the final result. Assume that there are K vertices in a plane graph G_p, represented as $v_i \in V$ for a vertex. Each edge is a pair of vertices, represented as $(v_i, v_j) \in E$, $\forall i < j$. Our goal is to represent the non-local relations among features and aggregate those non-local information, which reflects that the radiologists will leverage those non-local but easily-observed similar tissues for auxiliary observation. The first step is to indicate the importance of the vertex j's feature to vertex i's one. In image domain, the importance of one pixel relatively to another is usually measured by similarity [21]. Instead of relying on the weight matrix to obtain the importance, e.g., self-attention [25], in this paper, we propose to express the importance by the pixel-wise Euclidean distance in the feature space, which is inspired by Non-local Means (NLM) [26] for patch-wise non-local denoising. As a non-deep-learning method, NLM-based LDCT denoising method [27] has impressive adaptive denoising ability for complex noise level in real environments, which will also be desired for existing deep learning-based models. Formally, the similarity can be computed as follows

$$e_{ij} = \frac{\|v_i - v_j\|_2^2}{h_{ij}}, \tag{1}$$

where h_{ij} denotes the square root of the dimension of the v_i or v_j. In order to reduce computational complexity, we compute the e_{ij} for pixels $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is a $d \times d - 8$ (The total number of directly adjacent pixels for vertex i is 8) non-local neighborhood region of vertex i. Instead of leveraging all pixels in a CT slice according to the importance [21], we only select (K-1) nearest neighbour (namely K-NN) in feature space for information aggregation, which can avoid ineffective computation and also achieve the desired performance. In this

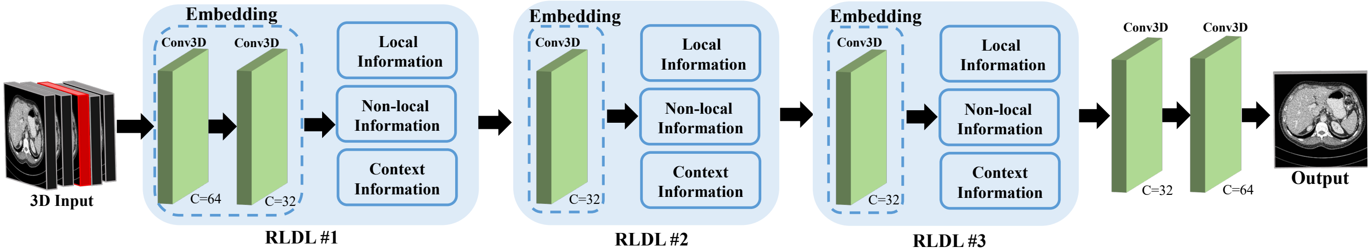


Fig. 3. The proposed stacked radiologist-inspired deep denoising networks. This model stacks three RIDnet modules. The last two convolution layers aim to generate a channel encoder-decoder structure as suggested in [16], [17].

paper, the K is set to 8 as discussed in supplement II. We then propose to normalize them across all choices of j such that the importances between vertices can be represented easily.

$$a_{ij} = softmax_j(e_{ij}) = \frac{exp(\frac{-\|v_i - v_j\|_2^2}{h_{ij}})}{\sum_{k=1}^{K-1} exp(-e_{ik})}. \tag{2}$$

One can observe the importance of neighbor vertices v_j's of v_i is represented by the probabilistic result a_{ij} induced from the distance of feature space. In this paper, we assign every edge (v_i, v_j) as a_{ij}. The next step is to aggregate those non-local information on this weighted graph. In order to achieve the specific-task fashion, inspired by the Edge-Conditioned Convolution (ECC) [28], the probability-based edge is introduced into ECC as follows:

$$\begin{aligned} \mathbf{s}_i^{N\_L} &= \frac{1}{K-1}\sum_{j=1}^{K-1}(F(a_{ij}, w)\mathbf{v}_{ij} + b) \\ &= \sum_{j=1}^{K-1}\frac{\Theta_{ij}\mathbf{v}_{ij}}{(K-1)} + b, \end{aligned} \tag{3}$$

where F^l denotes the output of a network parameterized by w^l which is used to dynamically produce the weight matrix $\Theta^l_{j,i}$ for different edge labels $z^{l,j\to i}$, and $\mathbf{b}^l$ is a learnable bias. The optimization details can be found in [28]. In our work, we adopt a multi-layer perception network for F^l,

4) Depth Graph Convolutional Module.: Instead of searching the non-local information in the intra-slice, depth graph convolutional module aims to explicitly obtain the useful similar information in the inter-slices, which also reflects the behavior of the radiologists, i.e., sliding the mouse to leverage front-and-rear slices that helps the denoising of the ROI with heavy noise. For every pixel in the feature space of a CT slice, let construct a depth graph $G_d = (V, E)$. Assume that there are M vertices in a depth graph G_d, represented as $v_i \in V$ for a vertex. Each edge is a pair of vertices, represented as $(v_i, v_j) \in E$, $\forall i < j$. Our goal is also to represent the context relations among features and aggregate them. We adopt the same aggregation method as described in *Plane Graph Convolutional Module*. In this paper, M is set to 3, as the number of slices for a 3D input is not large.

5) The layer of extracting local information.: In practice, the radiologist will firstly concentrate on the local region of the lesion. The local information thus is extremely important for the denoising of noisy region. As in previous studies [15], the 2D convolution operation with the 3×3 filter is utilized to extract the local information of the feature map that needs to be denoised.

6) The layer of feature fusion.: Here, the non-local information, local information and context information have been obtained. There is a troublesome issue, i.e., how do we combine them? It may be very easy for radiologists to handle this fusion of comprehensive information. In order to address this issue, we must firstly review the workflow of radiologists. Intuitively, the radiologist, compared with the context information, will pay more attention to the inner information (including non-local information and local information) of a slice that needs to be denoised. Furthermore, due to the difference of slice thickness for different body regions or imaging vendors, the usability of inter-slice information may be not stable, because if the thickness of a set of slices is very large, the relationship of inter-slice won't be very strong for learning. We thus consider the context information into an auxiliary information. Based on aforementioned discussions, the process of feature fusion can be represented as

$$\mathbf{x}''_i = \alpha \cdot Mean(\mathbf{p}'_{i,NL} + \mathbf{p}'_{i,L}) + (1-\alpha)\cdot \mathbf{p}'_{i,C}, \tag{4}$$

where $\mathbf{p}'_{i,NL}$, $\mathbf{p}'_{i,L}$), and $\mathbf{p}'_{i,C}$ denote the non-local information of $\mathbf{x}'_i$, the local information of $\mathbf{x}'_i$, and the context information of $\mathbf{x}'_i$, respectively. $\alpha \in [0,1]$ is a learnable parameter. The α is initialized to 0. $Mean$ operator denotes the pixel-wise average operation.

Finally, the result of fusion $\mathbf{x}''_i$ is proposed to replace the feature map in original input position (as shown in Figure 2), i.e., $\mathbf{X}'_i = concat(\mathbf{x}_{-s}, \mathbf{x}_{i-1}, \mathbf{x}''_i, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_s)$, due to the considerations of keeping original 3D shape and a similar shortcut idea [16]. As shown in Figure 3, the overall model stacks three RIDnet modules. The last two convolution layers aim to generate a channel encoder-decoder structure as suggested in [16], [17]

7) The overall framework and loss function.: The overall framework is based on GAN. Specifically, the proposed RIDnet model plays the role of generator G. The discriminator D follows the structure in [17]. The loss function of generator is composed of adversarial loss and perceptual loss as [17]

$$\min_{\theta_G} = -\mathbb{E}_{\mathbf{X}_i} D(G(\mathbf{X}_i)) + \lambda \cdot \mathbb{E}_{\mathbf{X}_i,\mathbf{y_i}}\|\phi(G(\mathbf{X}_i)) - \phi(\mathbf{y_i})\|_2^2. \tag{5}$$

As in previous studies, VGG19 [29] is adopted as the feature extractor ϕ. λ is a balance term that is set to 0.1 as suggested in [17]. θ_G denotes the parameters of the generator. The loss function of discriminator follows the wasserstein generative adversarial network based on gradient penalty [30]. The convergence process of the loss can be found in supplement III.

C. Datasets

1) The Training Set for Model Learning.: As previous studies [7], [9], [10], the public low/normal-dose dataset released from the 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge is used for training. This dataset includes 10 patient's abdominal examinations obtained on similar scanner models (Somatom Definition AS+, or Somatom Definition Flash operated in single-source mode, Siemens Healthcare, Forchheim, Germany). The normal-dose CT images are acquired under the settings of 120kV, 50mAs. Note that the low-dose counterparts are simulated to reach a noise level that corresponded to 25% by inserting the Poisson noise. More information about the dataset is available in [1]. We use 6 patient's CT images as the training set and 4 patients' CT images as validation set. To balance the learning efficiency and the memory consumption, every CT image with the size of 512×512 is randomly divided into non-overlapping 64×64 sub-patches. The total number of sub-patches is about 27K.

2) The Test Set for Double-Blind Scoring Experiments.: To adequately compare the denoising performance of our proposed model with that of other state-of-the-art models, a paired low/normal-dose clinical CT dataset with various examination regions, i.e., Low-Dose CT images and Projection Data (a.k.a., LDCT-PD) dataset [2], is used in this study. The LDCT-PD dataset was released by Mayo Clinic, including 99 non-contrast head CT scans acquired for the patients of acute cognitive or motor deficit, 100 low-dose non-contrast chest scans acquired to screen high-risk patients for pulmona

ry nodules, and 100 contrast-enhanced CT scans of the abdomen acquired to look for the patients of metastatic liver lesions. Similar to [7], the chest and abdomen CT images are selected for double-blind study. Every part consists of 67 patient's scans. The normal-dose chest and abdomen CT images are scanned under the settings of 120kV, 250mAs and 100kV,300mAs, repectively. The corresponding low-dose images are 10% and 25% of normal-dose. It should be noted that the LDCT-PD dataset provided a well-written clinical report, including the labeled locations of the lesion and the diagnosis types of the lesion. Based on this, we thus select a CT slice labeled with lesion and its adjacent slices (a lesion slice and its front 4 slices and rear 4 slices, totally 9 CT slices) to denote overall slices of a patient. We then can study the influence of different denoising results for the lesion.

There are two points that need to be noted. First, the imaging parameters, equipment vendors and acquired locations have differences between the training set and test set, which well reflects the complex environments in clinic. To obtain the competitive denoising performance, the deep learning-based models must have a good adaptive and generalized capacity. This is one of the motivations for the double-blind test that if the deep learning-based denoising models have the effectiveness in clinic, and which models have the best denoising ability under complex conditions. Second, the test images in our double-blind experiment refer to plenty of CT images with labeled lesions, which is very meaningful for the radiologists, allowing them to evaluate if the denoising results will influence the judgment of lesion type or image feature.

D. The details of double-blind reader study

The low-dose CT images of each selected patient are denoised by three deep learn ing-based models (MAP-NN, RedCNN, and our proposed RIDnet). For each patient, 5 sub-folders can be obtained, i.e., 3 sub-folders with different denoising results, LDCT sub-folder, and NDCT sub-folder. The sub-folders of denoising results are named randomly (such as measure1, measure2, and measure3). Finally, the total 134 folders are obtained for double-blind study. Three experienced radiologists (J. Shen, radiologist #1 with 23 years experience, D. Wu, radiologist #2 with 10 years experience, X. Pan, radiologist #3 with 14 years experience) participate in this experiment. For the standard of evaluation, we adopt the scheme of 4-point scale as in the previous study [3] in terms of image noise, structural fidelity and overall score. Specifically, 1 score denotes the quality of the CT image as unacceptable for clinical diagnosis. 2 score denotes the CT image can provide limited diagnostic information only. 3 score denotes that the CT image is acceptable and can provide the average diagnostic information. 4 score denotes the CT image has a good quality in terms of the accurate diagnosis and interpretation. Unlike any other systemic double-blind study in [3], we add the overall score as a part of the evaluations. As some results may have very good denoising performance in terms of the ROIs but be sub-optimal in terms of the non-ROIs, the radiologists thus can reflect the real image quality using the term of overall score.

1) Training Details.: For the details of preprocessing of training set, we normalize the data in CT domain (the provided dataset is the format of DICOM) into 0 to 1 using different window width and window level, according to different body regions. For the abdomen scans, the window width and window level are 400, 40, respectively. For the chest scans, the window width and window level are 1500, -600, respectively. The outputs of the model are re-normalized into corresponding range of CT domain to generate the DICOM files. For the details of model training, the tensorflow [3] is applied to construct the proposed RIDnet model. The Adam optimizer [31] is used to optimize the parameters of the model with the learning rate of exponential decay. As following, the framework of generative adversarial network [32], the initial learning rate is set to 1×10^{-4} for the generator and 4×10^{-4} for the discriminator as suggested in [4]. To balance the memory consumption and learning efficiency, the batch size is set to 32. The model is trained with 40 epochs and converges finally, using 4 Geforce 1080Ti GPUs. We chose

[1] https://www.aapm.org/GrandChallenge/LowDoseCT/#

[2] https://wik i.cancerimagingarchive.net/pages/viewpage.action?pageId =52758026

[3] www.tensorflow.org

TABLE I
THE DETAILS OF DEEP LEARNING-BASED METHODS FOR COMPARISON

Method	RED-CNN [TMI,2017]	MAP-NN [Nature MI,2019]	RIDnet (Ours)
Backbone	CNN	CNN	GCN+CNN
Imitating Radiologist	No	No	Yes
Traning Set	AAPM-Mayo	AAPM-Mayo	AAPM-Mayo

the model that has the best performance on validation set for comparison experiment.

III. Results

To compare the performance of representative deep lear ning-based methods, a double-blind study is carried out in this paper. The representative methods for comparison include RED-CNN [19], MAP-NN [7], and our proposed RIDnet model. The corresponding details can be found in Table 1. Using mean square error (MSE) as the loss function, RED-CNN is widely adopted due to its well noise suppression. Through systemic double-blind study, MAP-NN shows competitive denoising performance compared with commercial iterative reconstruction methods. It should be noted that our proposed model is the only one that tries considering the behavior of the radiologists. To be fair, all methods are trained on AAPM-Moyo dataset, which includes 10 patient's abdomen scans (paired normal-dose and low-dose (25% of normal-dose) CT images). We utilize a separate dataset, i.e., LDCT-PD dataset, to evaluate the denoising capacity of different methods in real-world complex environments. In LDCT-PD dataset, the selected test data totally includes 67 pairs of normal/low-dose (25% of normal-dose) abdomen scans and 67 pairs of normal/low-does (10% of normal-dose) chest scans. It can be easily noted that the dose level of low-dose chest scans in test set is extremely lower than that of the training one. These test models thus must have a adaptive and generalized capacity to handle this various environment, which we will analyze further later in this paper. Unlike previous double-blind studies, the denoising performance of different methods is completely evaluated on the region of lesion. We believe that this may be more valuable for clinical environments. Specifically, the abdomen and chest CT images have the lesion of metastatic liver and the lesion of pulmonary nodule, respectively. Three experienced radiologists from West China Hospital and West China No.4 Hospital participated in this study. They are named as "radiologist #1", "radiologist #2", and "radiologist #3", respectively. The standard of 4-point scale is used to evaluate the denoising performance in terms of noise, fidelity, and overall score. Briefly, the higher the score, the better. More details about the settings of double-blind study can be found in the section Method.

The analysis of the results for the scans of the chest. As illustrated in the first column of Figure 4, we can make four important observations. First, our proposed model significantly outperforms other methods in terms of noise, fidelity, and overall score, except for a slight improvement in two results (noise and fidelity) from radiologist #3. This shows the superiority of integrating the radiologist-inspired denoising mechanism. Second, we can notice that RED-CNN is significantly behind MAP-NN and our proposed mod -el. Although the framework of the MSE-guided (adopted by RED-CNN) would intuitively generate very smooth and little noisy results (see the Figure 6 and Figure 7), the radiologist may not favour this style that is much limited for the improvement of the diagnosis. Instead, the results of generative adversarial network (GAN) framework (adopted by MAP-NN and ours) make an obvious improvement compared with LDCT. Third, compared with MAP-NN, our proposed model still achieved better performance despite using the same framework. Taking a different approach, the structure of obeying the workflow of radiologists is introduced in our proposed model such as non-local and context information, which appears to be very effective. Fourth, the better results reflect that our proposed model has better adaptive and generalized capacity, because the dose level of the test set mismatches that of the training set, being very much lower. This property will take a strong potential in complex clinical environments, especially for various imaging parameters, equipment vendors and acquired locations.

The analysis of the results for the scans of the abdomen. The second column of Figure 4 shows the results of evaluation for the abdomen. One has the following observations: First, putting together the results of three metrics, our proposed model achieves the best performance, especially for the term of fidelity. This also benefits greatly from the superiority of the radiologist-inspired model. The proposed model leverages comprehensive information, which is intuitively useful for the structure preservation compared with single information (such as RED-CNN and MAP-NN). Second, as the lesion of metastatic liver is usually characterized as the closely black region, the fidelity of grayscale and shape thus is very important for the diagnosis. Interestingly, the fidelity evaluation of our proposed model significantly outperforms other models in all radiologist's results. This will contribute to the diagnosis of some unobservable lesions.

The analysis of the results higher than LDCT in all cases. The number of cases with scores higher than LDCT images is also important to indicate the overall performance of denoising results for all cases, except the average score of different methods for different body regions. As shown in the first column of Figure 5, we have some observations. First, in terms of results of radiologist #1 and #2, we can find that our proposed model achieves improvement for all cases basically regardless of the noise, the fidelity, and the overall score. Although MAP-NN achieves good performance for the noise and overall score (our proposed model still has a slight superiority), the improvement of fidelity can not cover all cases. Intuitively, it is very hard to balance the noise removal and structure fidelity. The results of MAP-NN clearly show this challenge. Instead, our proposed model enjoys extra non-local and context information, which shows that it is more suitable to handle this challenge by integrating radiologist-inspired workflow. Second, due to a possible preference among radiologists, the improvements do not cover all cases in the radiologist #3's results. However, our proposed model still significantly outperforms others in term of the fidelity.

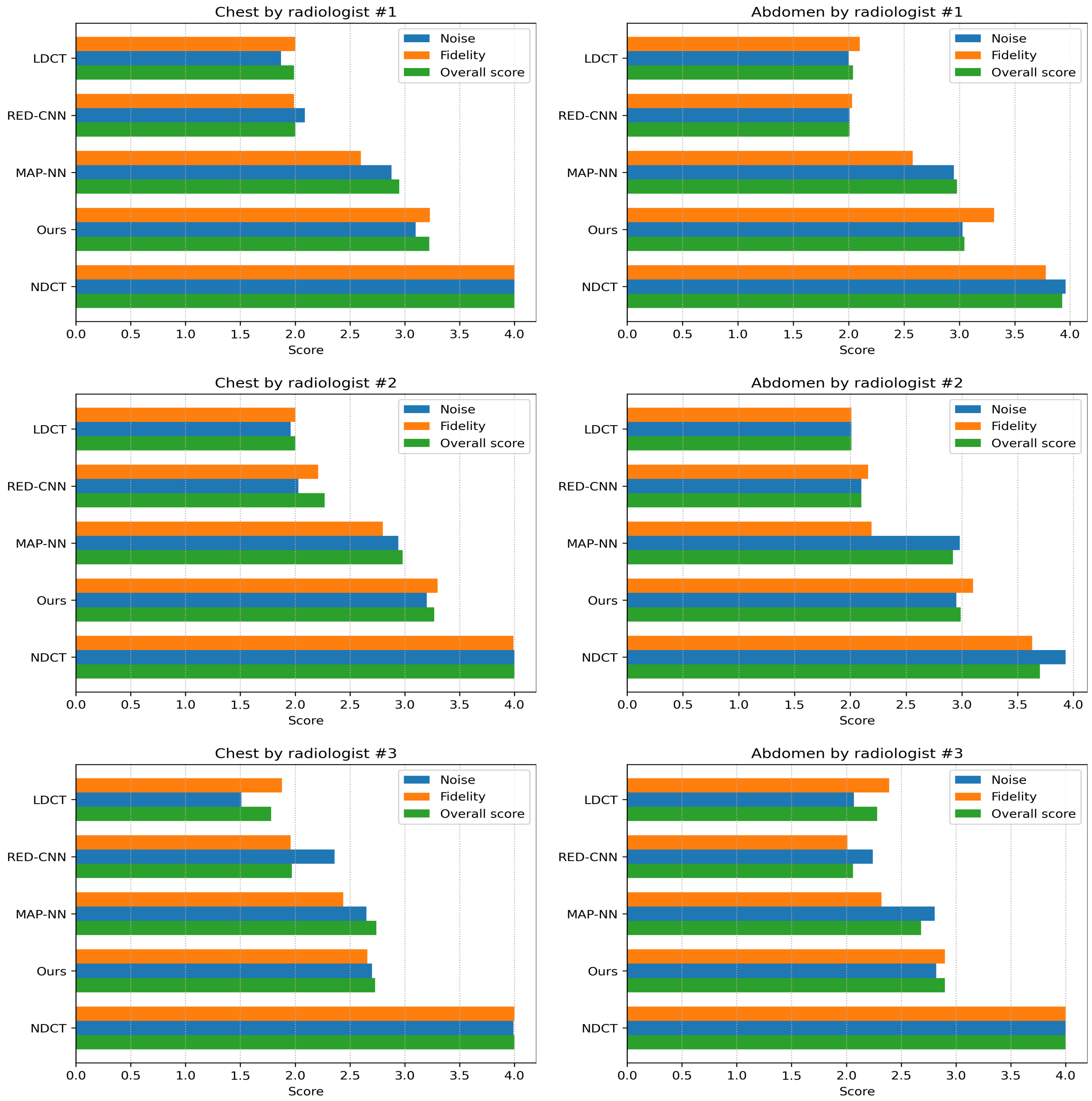


Fig. 4. The results of double-blind study by three radiologists aspect to the chest and the abdomen. For the noise, fidelity, and overall score, the higher the better. The full score is 4 for every evaluation standard.

The analysis of the results equal to NDCT in all cases. We are interested in how many denoising cases achieved the level of NDCT images, which can further represent the denoising capacity. Our proposed model achieve the best performance by analyzing all radiologist's results as shown in the second column of Figure 5. Interestingly, in the radiologist #3's results, our proposed model is the only one that can achieve the level of NDCT images for the fidelity. In summary, benefiting from the integration of radiologist-inspired behavior, our proposed model has the greatest potential to achieve the quality of NDCT images.

The analysis of visual results for zoomed-in subtle structure and lesion. *Abdomen:* As shown in Figure 6, we can make some important observations. First, the green box shows the comparison of zoomed-in subtle structure, we can find these structures (within red circles) nearly disappear for RED-CNN, due to the easily observed over-smoothness. Similarly, MAP-NN suffers from the same problem and also has a slight over-smoothness. Instead, our proposed model greatly preserves the subtle structure and generates the texture closest

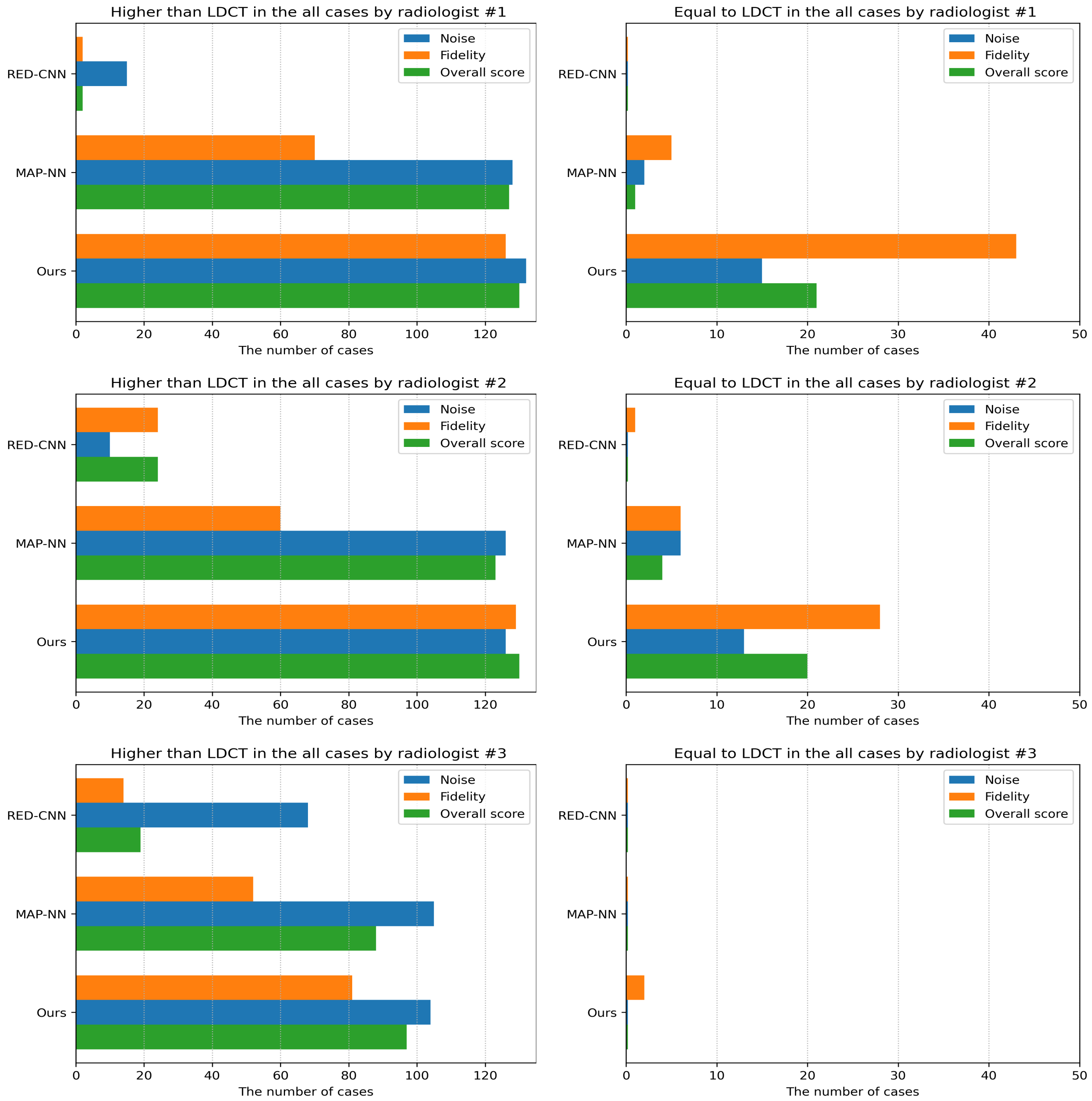
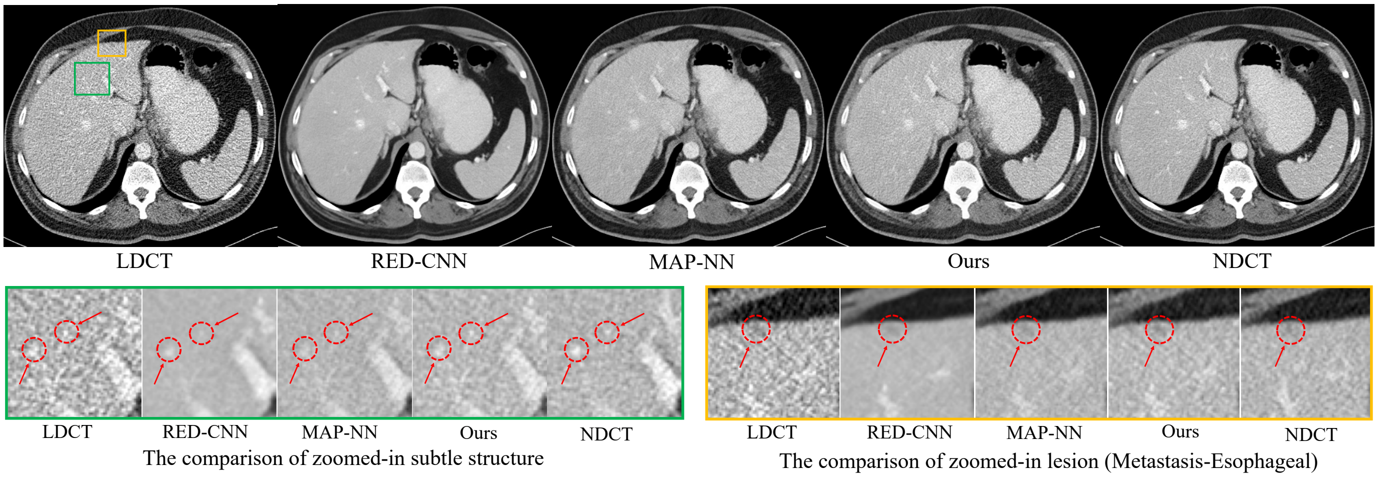


Fig. 5. The results of double-blind study by three radiologists aspect to all cases.

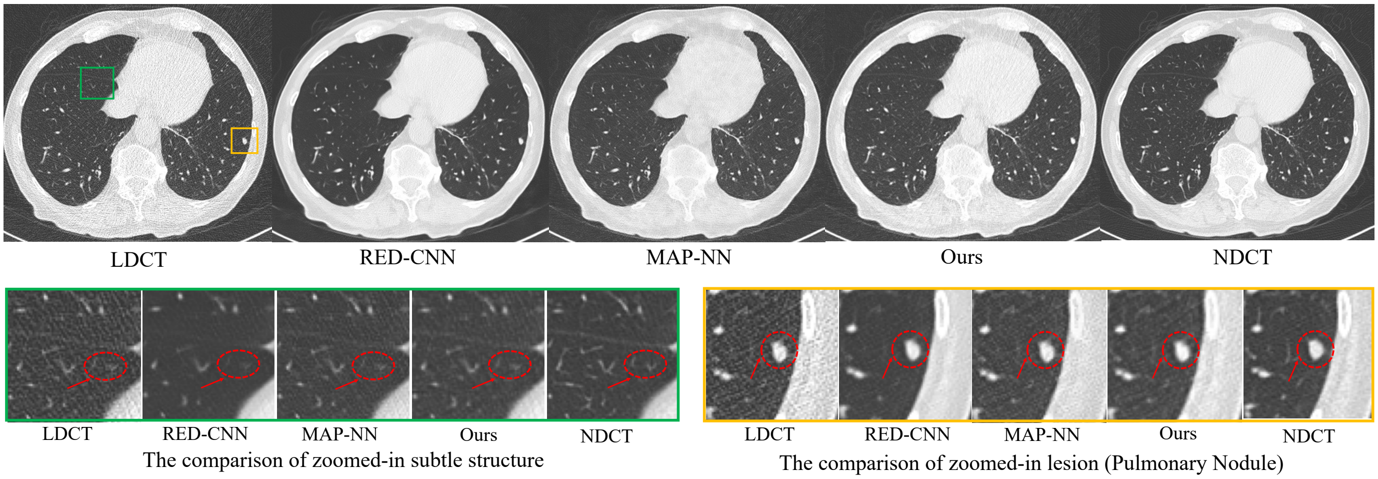
to NDCT. Both RED-CNN and MAP-NN only leverage the local information by the CNN, so it is difficult to balance the local detail and noise removal with limited information. Our proposed model adopts the framework of radiologist-inspired comprehensive information such that extra information can be used as a supplement to produce the optimal results regardless of structure and texture. Second, as illustrate in the red circle of yellow box in Figure 6, our proposed model has the most obvious observation for the lesion (diagnosed as Metastasis-Esophageal), especially for the level of the grayscale. However, the lesion in RED-CNN and MAP-NN becomes very fuzzy. The superiority of lesion region further proves the effect of radiologist-inspired framework. *Chest:* As shown in the yellow box of Figure 7, we can find that all models achieve the preservation of lesion (within red circle). However, MAP-NN has more easily observed noise points compared with RED-CNN and our proposed model. RED-CNN losses the subtle structure basically as shown in red circle of green box. In summary, our proposed model has the most impressive visual performance, which naturally obtains the best performance of double-blind study as reported in Figure 4 and 5.



The comparison of zoomed-in subtle structure

The comparison of zoomed-in lesion (Metastasis-Esophageal)

Fig. 6. The visual comparisons for zoomed-in subtle structure and lesion, in term of an example of the abdomen. The green box and yellow box are the region of zoomed-in subtle structure and the region of zoomed-in lesion, respectively. The red circles and arrows are the suggested region for comparison.



The comparison of zoomed-in subtle structure

The comparison of zoomed-in lesion (Pulmonary Nodule)

Fig. 7. The visual comparisons for zoomed-in subtle structure and lesion, in terms of an example of the chest. The green box and yellow box are the region of zoomed-in subtle structure and the region of zoomed-in lesion, respectively. The red circles and arrows are the suggested regions for comparison.

IV. Discussion

Compared with existing deep learning-based LDCT denoising methods, the superiority of our proposed method can be summarized in four parts. First, our proposed model is designed to imitate the behavior of radiologists and their workflow deeply, by comprehensively integrating the local, non-local, and context information for LDCT image denoising. By contrast, existing models lack the introduction of radiologist's behavior or workflow. Second, via the radiologist-inspired workflow, the proposed model shows the best adaptive and generalized capacity despite complex clinical environments (As shown in Figure 4, our proposed model has the most competitive performance under the challenge of dose mismatch between training set and test set). Third, how to balance the detail preservation and the level of noise removal is always a dilemma for existing deep learning-based models. However, our proposed model not only preserves the subtle structure and the lesion but also achieves the closest texture to NDCT (The texture can be regarded as the level of noise removal. If the level of noise removal is very high, the texture of denoising result will be very smooth). This adaptive capacity may benefit from the introduction of graph convolution, which also can be regarded as learnable Non-local Means [33] (that has perfect adaptive ability as a non-deep-learning method). Fourth, our double-blind study is completely based on the evaluations of lesion region. We believe that this is more valuable for clinical purposes.

To the best of our knowledge, we are the first to attempt to integrate the radiologist-inspired workflow into a deep neural network. In the future, we believe that more potential radiologist's behaviors should be considered to take further improvements such as generalized capacity and robustness. In addition, in real clinical environments, various scan conditions (such as the differences of vendor, reconstruction type, and imaging parameters) will inevitably present a huge challenge for LDCT images denoising. We must be careful about this and design some methods to address it. In practice, we highly recommend using the double-blind reader study to evaluate different denoising methods, because existing objective metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural

Similarity Index Measure (SSIM) can not fully reflect the quality of denoising results (For example, the over-smooth result usually has a high PSNR score). However, the double-blind study will take a lot of time for radiologists. We thus need some better objective metrics, especially for the region of lesion.

V. Conclusion

In this study, we propose a novel deep learning model named radiologist-inspired deep denoising network (RIDnet) to imitate the workflow of a radiologist reading LDCT images. A double-blind reader study on a public clinical dataset shows the effectiveness of proposed model. As a physicians-inspired model, RIDnet gives a new research road map that takes into account the behavior of physicians when designing decision support tools for assisting clinical diagnosis.

References

[1] T. M. Buzug, "Computed tomography," in *Springer Handbook of Medical Technology*. Springer, 2011, pp. 311–342. I

[2] C. Lischer, U. Walliser, P. Witzmann, M. W. Eser, and S. Ohlerth, "Fracture of the paracondylar process in four horses: advantages of ct imaging," *Equine veterinary J.*, vol. 37, no. 5, pp. 483–487, 2005. I

[3] D. Kim, S. Park, J. H. Lee, Y. Y. Jeong, and S. Jon, "Antibiofouling polymer-coated gold nanoparticles as a contrast agent for in vivo x-ray computed tomography imaging," *Journal of the American Chemical Society*, vol. 129, no. 24, pp. 7661–7665, 2007. I, II-D, II-D

[4] V. Rosso, N. Belcari, M. G. Bisogni, C. Carpentieri, A. Del Guerra, P. Delogu, G. Mettivier, M. Montesi, D. Panetta, M. Quattrocchi *et al.*, "Preliminary study of the advantages of x-ray energy selection in ct imaging," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 572, no. 1, pp. 270–273, 2007. I, II-D1

[5] M. S. Pearce, J. A. Salotti, M. P. Little, K. McHugh, C. Lee, K. P. Kim, N. L. Howe, C. M. Ronckers, P. Rajaraman, A. W. Craft, L. Parker, and A. Berrington de González, "Radiation exposure from ct scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study," *The Lancet*, vol. 380, no. 9840, pp. 499 – 505, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673612608150 I

[6] S. M. H. Tabatabaei, H. Talari, A. Gholamrezanezhad, B. Farhood, H. Rahimi, R. Razzaghi, N. Mehri, and H. Rajebi, "A low-dose chest ct protocol for the diagnosis of covid-19 pneumonia: a prospective study," *Emergency Radiology*, vol. 27, no. 6, pp. 607–615, 2020. I

[7] H. Shan, A. Padole, F. Homayounieh, U. Kruger, R. D. Khera, C. Nitiwarangkul, M. K. Kalra, and G. Wang, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose ct image reconstruction," *Nature Machine Intelligence*, vol. 1, no. 6, pp. 269–276, 2019. I, I, II-C1, II-C2, III

[8] Y. Liu, J. Ma, Y. Fan, and Z. Liang, "Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction," *Physics in Medicine & Biology*, vol. 57, no. 23, p. 7923, 2012. I

[9] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *TMI*, vol. 37, no. 6, pp. 1348–1357, 2018. I, II-C1

[10] K. Choi, J. S. Lim, and S. Kim, "Statnet: Statistical image restoration for low-dose ct using deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1137–1150, 2020. I, II-C1

[11] J. Wang, H. Lu, T. Li, and Z. Liang, "Sinogram noise reduction for low-dose ct by statistics-based nonlinear filters," in *MIIP*, vol. 5747. International Society for Optics and Photonics, 2005, pp. 2058–2066. I

[12] A. Manduca, L. Yu, J. D. Trzasko, N. Khaylova, J. M. Kofler, C. M. McCollough, and J. G. Fletcher, "Projection space denoising with bilateral filtering and ct noise modeling for dose reduction in ct," *Medical physics*, vol. 36, no. 11, pp. 4911–4919, 2009. I

[13] M. Balda, J. Hornegger, and B. Heismann, "Ray contribution masks for structure adaptive sinogram filtering," *IEEE TMI*, vol. 31, no. 6, pp. 1228–1239, 2012. I

[14] Y. Zhang, X. Mou, G. Wang, and H. Yu, "Tensor-based dictionary learning for spectral ct reconstruction," *IEEE TMI*, vol. 36, no. 1, pp. 142–154, 2016. I

[15] K. Chen, X. Pu, Y. Ren, H. Qiu, H. Li, and J. Sun, "Low-dose ct image blind denoising with graph convolutional networks," in *Neural Information Processing*, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham: Springer International Publishing, 2020, pp. 423–435. I, II-B5

[16] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2289–2301, 2020. I, 2, 3, II-B6, II-B6

[17] H. Shan, Y. Zhang, Q. Yang, U. Kruger, M. K. Kalra, L. Sun, W. Cong, and G. Wang, "3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network," *TMI*, vol. 37, no. 6, pp. 1522–1534, 2018. I, I, 3, II-B6, II-B7, II-B7, II-B7

[18] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose ct via convolutional neural network," *Biomedical optics express*, vol. 8, no. 2, pp. 679–694, 2017. I

[19] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *TMI*, vol. 36, no. 12, pp. 2524–2535, 2017. I, III

[20] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *TMI*, vol. 37, no. 6, pp. 1348–1357, 2018. I

[21] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging*, 2020. I, I, II-A, II-B3, II-B3

[22] L. Huang, H. Jiang, S. Li, Z. Bai, and J. Zhang, "Two stage residual cnn for texture denoising and structure enhancement on low dose ct image," *Computer Methods and Programs in Biomedicine*, vol. 184, p. 105115, 2020. I

[23] I. Elezi, "Exploiting contextual information with deep neural networks," 2020. I

[24] D. Valsesia, G. Fracastoro, and E. Magli, "Deep graph-convolutional image denoising," *arXiv preprint arXiv:1907.08448*, 2019. II-B1, II-B1

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. II-B3

[26] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, ser. CVPR '05. USA: IEEE Computer Society, 2005, p. 60–65. II-B3

[27] Z. Li, L. Yu, J. D. Trzasko, D. S. Lake, D. J. Blezek, J. G. Fletcher, C. H. McCollough, and A. Manduca, "Adaptive nonlocal means filtering based on local noise level for ct denoising," *Medical physics*, vol. 41, no. 1, p. 011908, 2014. II-B3

[28] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *CVPR*, 2017, pp. 3693–3702. II-B3, II-B3

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. II-B7

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017. II-B7

[31] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th IWQoS)*. IEEE, 2018, pp. 1–2. II-D1

[32] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019. II-D1

[33] Z. Li, L. Yu, J. D. Trzasko, D. S. Lake, D. J. Blezek, J. G. Fletcher, C. H. McCollough, and A. Manduca, "Adaptive nonlocal means filtering based on local noise level for ct denoising," *Med Phys*, vol. 41, no. 1, p. 011908, 2014. IV

APPENDIX A
THE OPTIMAL NUMBER OF RIDNET

It is very important to determine how many RIDnets are optimal for denoising. To this end, we construct 5 models. They have 1, 2, 3, 4 and 5 RIDnets, respectively. For evaluation metrics, we use state-of-the-art Radiomics features, including the contrast and the dissimilarity. The contrast reflects the clarity of the image and the local feature of the texture. The correlation measures the similarity of local gray level values. We use the absolute loss between the contrast/correlation value of NDCT images and that of denoising results to evaluate the performance. For the contrast and correlation losses, the lower the better.

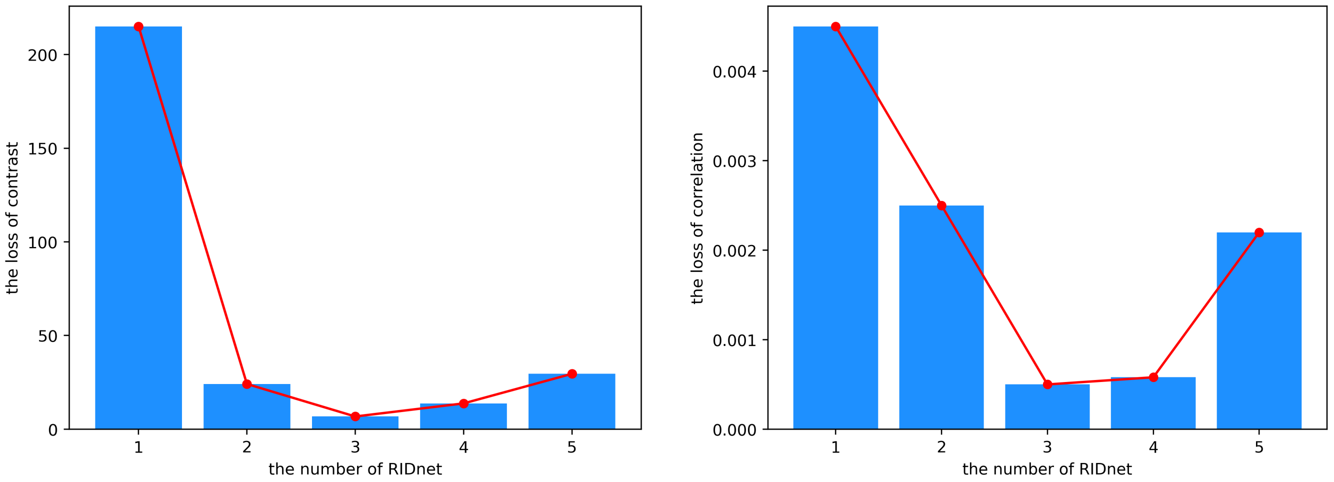
As illustrated in Figure 8, we can find that the model with 3 RIDnets achieves the best performance compared with others. Thus, in this paper, the adopted model is based on three RIDnets, as shown in Figure 3.

APPENDIX B
THE OPTIMAL NUMBER OF K

In this paper, K denotes the number of non-local information for a pixel. As shown in Figure 9 (a) and (b), we can find that the contrast loss decreases gradually with the increasing of K. The non-local information shows the effectiveness for model. Meanwhile, too much non-local information may not be useful. The losses of contrast and correlation generally increases when the number of K is greater than 8. To this end, the number of K is set to 8.

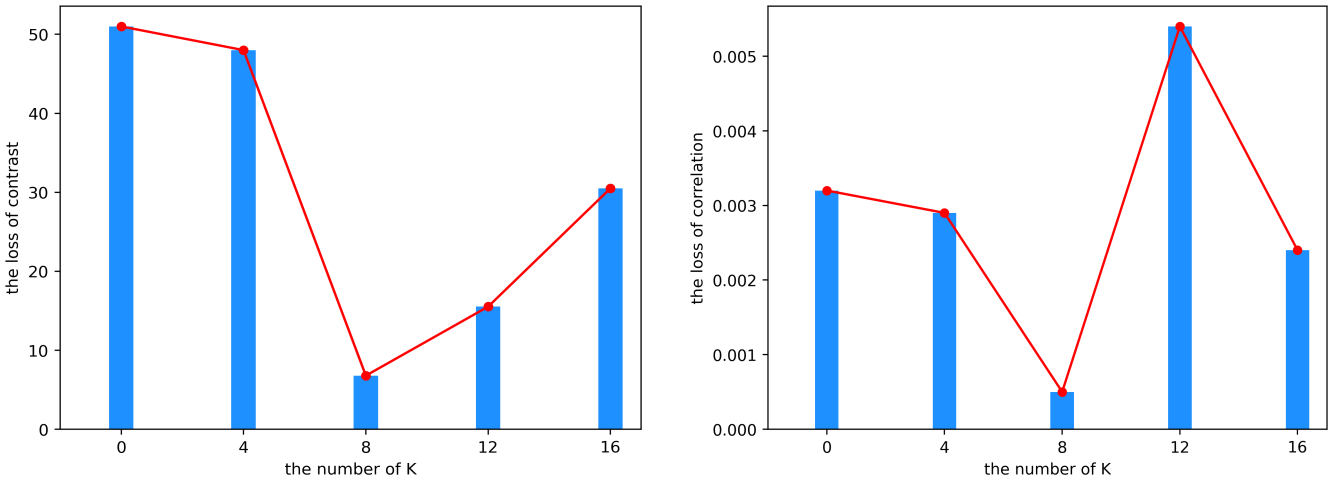
APPENDIX C
THE PROCESS OF MSE LOSS AND PERCEPTUAL LOSS WITH TRAINING

As illustrated in Figure 10(a), we can observe that the MSE loss converges gradually. The perceptual loss converges oscillatingly, as shown in Figure 10(b).



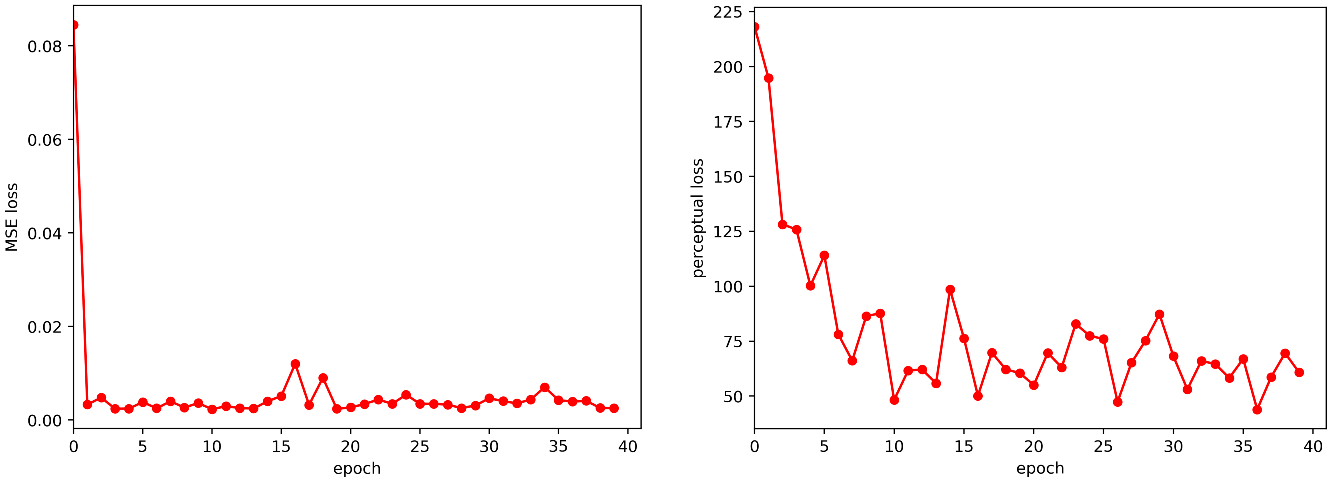
(a) (b)

Fig. 8. The results of optimal number of RIDnet. (a) The relation between contrast loss and the number of RIDnets. (b) The relation between correlation loss and the number of RIDnets.



(a) (b)

Fig. 9. The results of optimal K. (a) The relation between contrast loss and the number of K. (b) The relation between correlation loss and the number of K.



(a) (b)

Fig. 10. Analysis of model training(a) The process of MSE loss convergence (b) The process of perceptual loss convergence.