Two-Stage Convolutional Neural Network for Medical Noise Removal via Image Decomposition

Yi Chang, *Student Member, IEEE,* Luxin Yan, *Member, IEEE,* Meiya Chen, Houzhang Fang, and Sheng Zhong

***Abstract*—Most of existing medical image denoising methods focus on estimating either the image or the residual noise. Moreover, they are usually designed for one specific noise with a strong assumption of the noise distribution. However, not only the *random independent* Gaussian or speckle noise but also the *structurally correlated* ring or stripe noise, are ubiquitous in various medical imaging instruments. Explicitly modeling the distributions of these complex noises in the medical image is extremely hard. They cannot be accurately held by the Gaussian or mixture of Gaussian model. To overcome the two drawbacks, in this work, we propose to treat the image and noise components equally and convert the image denoising task into an image decomposition problem naturally. More precisely, we present a two-stage deep convolutional neural network (CNN) to model both the noise and the medical image simultaneously. On the one hand, we utilize both the image and noise to separate them better. On the other hand, the noise sub-network serves as a noise estimator which guides the image sub-network with sufficient information about the noise, thus we could easily handle different noise distributions and noise levels. To better cope with the gradient vanishing problem in this very deep network, we introduce both the short-term and long-term connections in the network which could promote the information propagation between different layers efficiently. Extensive experiments have been performed on several kinds of medical noise images, such as the computed tomography and ultrasound images, and the proposed method has consistently outperformed state-of-the-art denoising methods.**

***Index Terms*—Image denoising, image decomposition, convolutional neural network, medical image.**

I. Introduction

THE medical noises would obviously increase the uncertainties in the measurement procedures, and degrade the quality of the images seriously, which make them diagnostically unusable. Numerous image denoising methods have been proposed in the past decades. In this work, from a more general perspective, we define the noise (More precisely, we could name the noise here as the artifacts) as anything that is not expected to be presented in the medical images. The noises have different appearances in different imaging instruments and can be broadly classified into two categories: the independent random noise and the structurally correlated/fixed pattern noise. In this work, for the random noise, we mainly focus on the additive Gaussian random noise in computed tomography (CT) image and the multiplicative speckle noise in the ultrasound image. For the structural noise, we mainly focus on the line pattern stripe noise in the scanning electron microscope (SEM) image and the circle pattern ring noise in CT image, as shown in Fig. 1. Even worse, the mixture of noise makes the problem more intractable. The goal of this work is to suppress all these artifacts or the mixture of them via a unified image processing method.

To date, a variety of filter based medical noise removal methods [1]–[4] have been proposed such as directional filter [5], wavelet [6], [7]. Over the past decade, the sparse representation based optimization methods have received significant attention, which assumes that the images naturally have the sparsity property in a particular transformed domain, such as the total variational [8], [9], the dictionary learning [10]–[13] to name a few. When the non-local self-similarity meets the sparse representation, such as the group sparsity [14], [15] and low-rank representation method [16]–[18], the combination further boosts the medical denoising performance.

However, there are three main limitations of the optimization based methods. First of all, the optimization methods need explicitly model the statistical distribution of the noise. For the real-world medical images, where the noises are much more complex and vary with different appearances such as the structural noise or mixed noise, it is very hard to figure out the corresponding mathematical formulations precisely. The approximations via mixture of Gaussian (MoG) or Laplacian also fail to model the structural noise. Second, most of the previous optimization based methods mainly utilize a predefined prior. Such a hand-crafted prior is definitely not suitable for multi-modality medical images. Last but not least, the computational load is heavy due to the iteration and complex operation, which limits their potential for real-time application.

Recently, CNN has been widely used for low-level image tasks, such as image denoising [19], deblurring [20], and super-resolution [21]. The CNN does not need to model the distribution of the noise explicitly. Instead, for any arbitrary ‘noise’, the network can implicitly approximate them guaranteed by the universal approximation theory [22]. Moreover, the learned prior from the training pairs make the network more adaptive for specific images. At last, due to the simple operation of the network, its forward process is extremely fast which makes it quite suitable for real-time application. These essential advantages in noise modeling, data prior, and running

This work was supported in part by the projects of the National Natural Science Foundation of China under Grants No. 61571207, 61433007, and 41501371, and in part by the project of the Hubei Provincial Natural Science Foundation of China under Grants No. 2018CFA089.

Y. Chang, L. Yan, M. Chen and S. Zhong are with the National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China. Fax: 86-27-98753594; Tel: 86-27-87541761; e-mail:{yichang, yanluxin, my_chen, zhongsheng}@hust.edu.cn;

H. Fang is with School of Software, Xidian University, Xian, Shaanxi 710071, China (e-mail: houzhangfang@xidian.edu.cn)

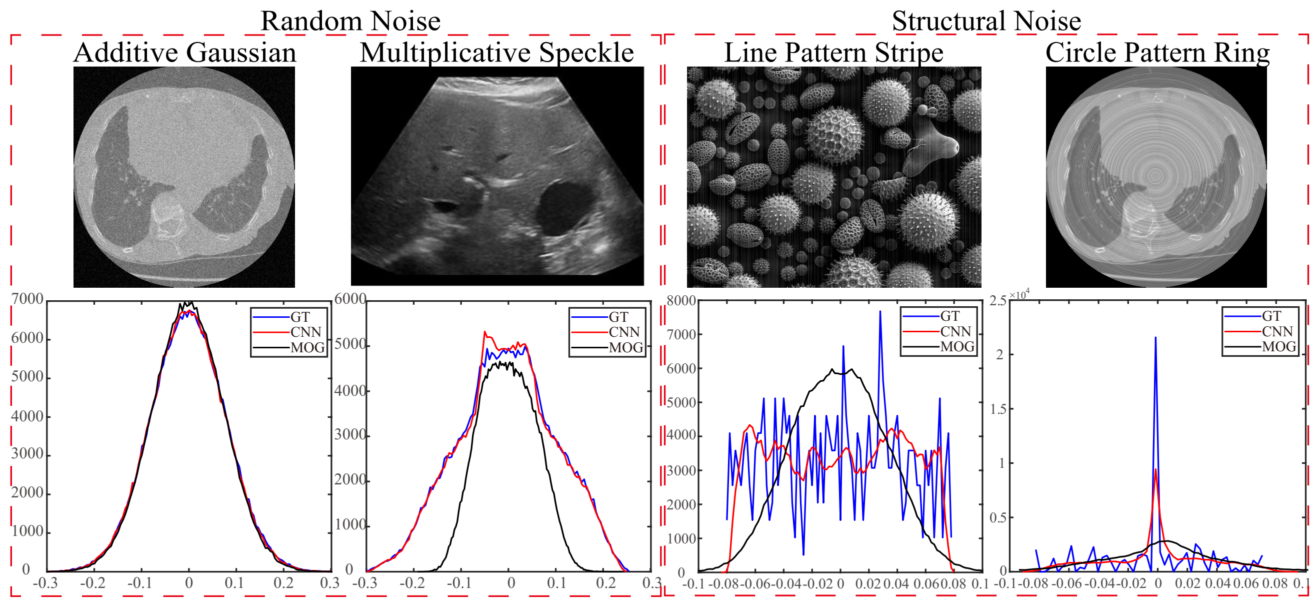


Figure 1. The advantageous of the CNN over the statistical based methods for modeling various kinds of medical noises. Representative medical noises and their statistical distributions are shown in the first and second row, respectively.

time make CNN a powerful tool for image denoising.

There have been emerged several pioneer deep learning (DL) works for medical image restoration [23]–[31]. The FBP-ConvNet [23] combined filtered back projection (FBP) with a multiresolution CNN for X-ray CT reconstruction. In [24], Chen *et al*. proposed a deep encoder-decoder convolutional neural network (RED-CNN) for low-dose CT imaging. The denoising autoencoder (DAE) was introduced to remove the Poisson and Gaussian random noise in the medical images [26]. Zhang *et al*. [28] developed a CNN based metal artifacts removal framework, which fused both the information from the original and corrected images to suppress artifacts in the X-ray CT images. Pham *et al*. [29] introduced the three-dimensional (3D) CNN for brain MRI image super-resolution with the help of patches of other HR brain images. Although the DL methods have achieved encouraging denoising performance, they still face some problems when dealing with various kinds of noises. *All of them focus on either estimating the image or the noise, whereas few of them take into account the characteristic of both the image and noise. We argue that both of them are important and beneficial to the denoising results.*

In this work, we address the noise removal problem from an image decomposition perspective, in which both the multi-modality image and noise are represented by different sub-networks. We model both the noise and image simultaneously within a cascaded CNN. In the first and second stage, we estimate the noise and image component respectively. The estimated noise in the first stage is further fed to the second sub-network along with the noisy image, which works as a conditional map to guide the attention of the second sub-network. These tasks are learned end-to-end by cascading two similar CNNs, and no hand-crafted modules are required.

As for the training, to avoid gradient vanishing problem, we introduce both the short-term and long-term connections for better feature propagation and reuse. Concretely, the short-term connections generate from the residual blocks [32], and the long-term connections generate from several skip connections. We show that the short-term and long-term connections jointly make the training of the deeper network easier and more effective with better restoration performance. To alleviate the issue of limited training samples, we firstly pre-train our model on the natural images and then fine-tune it on the corresponding limited medical images. The contributions of the proposed work are summarized as follows:

- We utilize the characteristics of both the noise and image component simultaneously from the image decomposition perspective. Thus, we propose a two-stage CNN model to restore the desired images by leveraging the predicted noise map, such that our model is robustness to different noise categories and noise levels.
- Our method can handle various medical noises with fast testing speed and better performance over the state-of-the-art methods. Extensive experimental results on different medical datasets verify the effectiveness and efficiency of the proposed method.

The remainder of this paper is organized as follows. In Section II, we analyze why we use CNN for the medical image and noise modeling, and present our two-stage very deep model. Experimental results and discussion are reported in Section III. Finally, we conclude the paper in Section IV.

II. Two-Stage Convolutional Neural Network For Medical Noise Removal

A. Preliminary

1) Image Decomposition: For medical images, the noises mainly contain the random and structural ones, as shown in

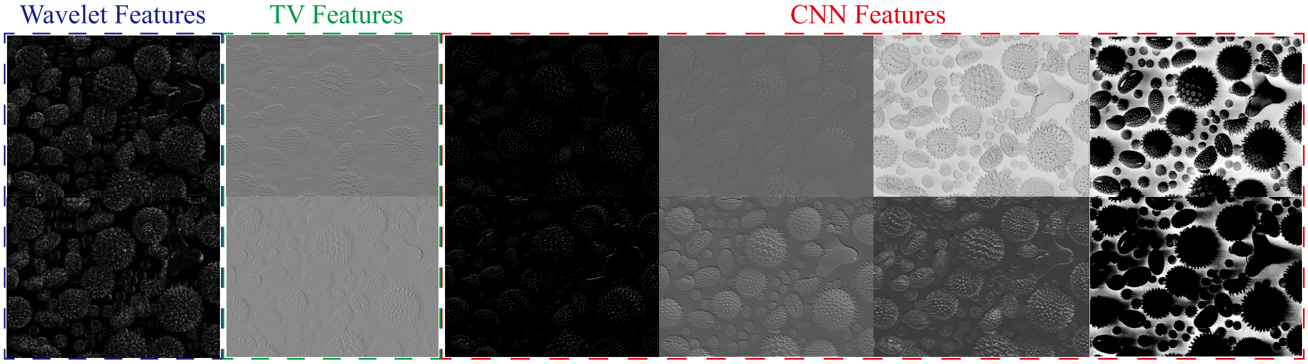


Figure 2. The advantageous of the CNN over the hand-crafted priors for modeling medical image structure. We visualize the features extracted by CNN and the hand-crafted methods. The TV and wavelet method consistently utilize the same features, while the CNN could extract the hierarchical features.

Fig. 1. In this paper, we assume that the degradation can be formulated mathematically as follow:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{X}$ is the ground truth image, $\boldsymbol{Y}$ is the input noise image, $\boldsymbol{N}$ is the noise. The goal of the image decomposition is to estimate both the clear image $\boldsymbol{X}$ and the noise $\boldsymbol{N}$ simultaneously from the degraded image $\boldsymbol{Y}$.

2) CNN: Assuming there are D layers in the designed network, for a given sample $\boldsymbol{Y} \in \mathbb{R}^{R\times C\times B}$, the output of the first layer is $\boldsymbol{X}^{(1)} = S(\boldsymbol{W}^{(1)} \otimes \boldsymbol{Y} + \boldsymbol{P}^{(1)}) \in \mathbb{R}^{R\times C\times B_1}$, where $\boldsymbol{W}^{(1)}$ is the projection matrix to be learned from the first layer, $\boldsymbol{P}^{(1)}$ is the bias vector, $\otimes$ is the convolutional operator, B_1 is the channel number of the first layer, and $S : \mathbb{R} \mapsto \mathbb{R}$ is the nonlinear activation function which handles each pixel individually, such as the *sigmoid* or rectified linear unit (*RELU*). Next, the output of the first layer $\boldsymbol{X}^{(1)}$ is treated as the input of the second layer. Consequently, the output of the d-th layer can be expressed as:

$$\boldsymbol{X}^{(d)} = S(\boldsymbol{W}^{(d)} \otimes \boldsymbol{X}^{(d-1)} + \boldsymbol{P}^{(d)}) \in R^{R\times C\times B_d}. \tag{2}$$

It is easy to understand that the forward procedure (namely the Eq. 2) is to extract the features from the input data in a hierarchy manner. To learn the parameters $\{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, ..., \boldsymbol{W}^{(D)}\}$, the back propagation is applied to solve the following problem:

$$\boldsymbol{J}^{I+N}_{Recon} = \frac{1}{2}\left\|\mathcal{F}_I([\boldsymbol{Y}, \hat{\boldsymbol{N}}]) - \boldsymbol{X}\right\|^2 + \frac{1}{2}\|\mathcal{F}_N(\boldsymbol{Y}) - \boldsymbol{N}\|^2, \tag{3}$$

where $\mathcal{F}_I$ and $\mathcal{F}_N$ are the composite network mapping functions for the image and noise, respectively. In the next subsections, we will analyze why we choose CNN to model both the image and noise component. Then, we will present our two-stage network in detail.

B. Why CNN for Noise Modeling: Statistical Analysis

In conventional methods, they always assume the distribution of the noise to be Gaussian or a mixture of Gaussian. However, the noise characteristic is complex in the medical image [33], which is hard to provide the concrete expression explicitly. Moreover, the structural noise with correlation makes the problem harder.

To illustrate this problem, we plot the statistical distribution of four representative noises in Fig. 1. we generate various noise images in the first row. For example, we generate the Gaussian noise by the Matlab function sigma_gau*randn(size($\boldsymbol{X}$)), the speckle noise via imnoise($\boldsymbol{X}$, ‘speckle’, var_speckle), the stripe noise by the random lines with different intensity, and the ring noise by the Bresenham circle [34]. The histograms (horizontal axis means the normalized intensity, vertical axis means the number of the corresponding intensity values.) of both the ground truth (blue curve), a mixture of Gaussian model [35] (black curve), and the proposed method (red curve) is shown in the second row.

We have three main observations here. First, the distributions of different noises vary from the random noise to the structural noise. It is extremely hard to explicitly fit the distribution of structural noise. That is the main reason why the structural noise is harder to be removed. Second, for the random noises, the distribution estimated by our method could perfectly match that of the ground truth, while the MoG method fails to accommodate the multiplicative speckle noise well. Third, for the structural noises, although the estimated distributions of our CNN method are not exactly matching to the original ones, the CNN provides a satisfactory approximation to the original ones. This powerful fitting ability of the deep model can be well explained by its universal approximation for arbitrary signal [22]. These observations motivate us to model the medical images with the CNN model.

C. Why CNN for Image Modeling: Three Explanations

In this section, we will illustrate why the CNN based model performs better than previous methods in three aspects: the intuitive, the mathematical, and the practical result aspect.

1) From Intuitive Perspective: The start point between the conventional methods (filter and optimization methods) and CNN method is opposite. The conventional methods treat the denoising task as an ill-posed problem by obtaining the desired solution from the degraded input. However, there exist infinite solutions for the inverse process.

On the contrary, the learning-based CNN methods place emphasis on the forward process: the preparation of the clean/degraded pairs for training. The CNN model then is trained to fit the degradation to clean mapping, which can be well addressed by the back propagation algorithms, such as the ADAM [36]. Thus, the preparation for the training pairs is a

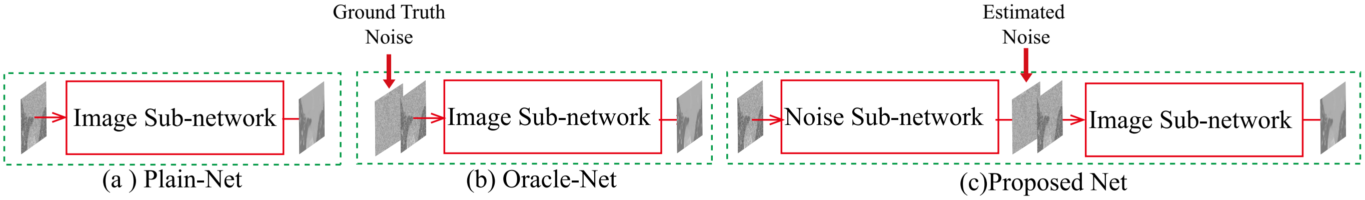


Figure 3. Illustration of the different networks. (a) Plain-Net. (b) Oracle-Net. (c) Proposed network. Compared with plain-Net, the oracle-Net has an additional ground truth noise map as the conditional input. Compared with oracle-Net, the proposed network learns the noise map via a sub-network in the first stage.

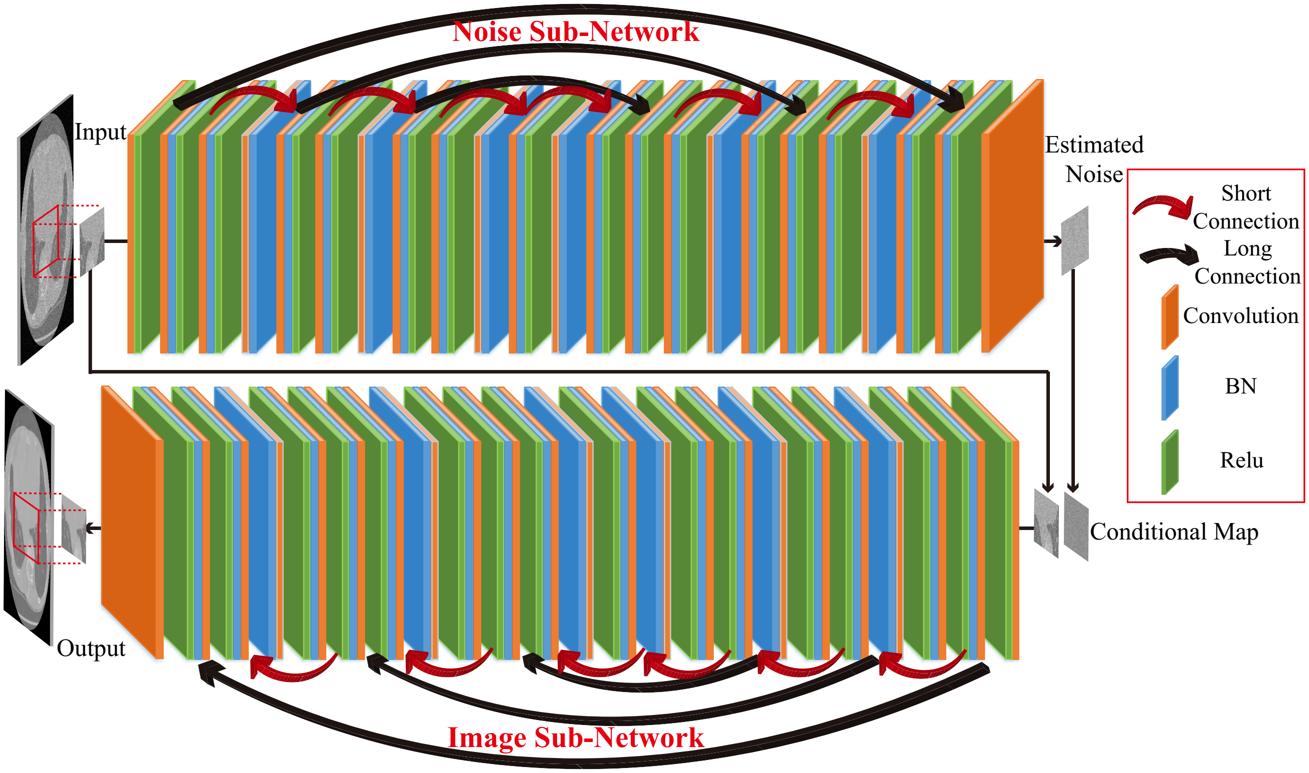


Figure 4. The architecture of the two-stage deep model. Two CNN sub-networks with similar structure are used for noise prediction and image reconstruction, respectively. The image sub-network takes the output of noise sub-network as a conditional map. In other words, the estimated noise map via the first noise sub-network serves as an guidance by indicating the image sub-network about the noise distribution and level for better restoration.

key factor in the success of the medical image noise removal. The CNN methods transfer the difficulty from solving the ill-posed problem (backward process) to the training data preparation (forward process. Normally, we only need to generate the noisy image from the ground truth), which significantly reduces the difficulty of the problem.

2) From Mathematical Perspective: For the filtering based methods, the solution can be roughly expressed as:

$$\hat{\boldsymbol{X}} = \Gamma(\phi(\boldsymbol{Y})), \tag{4}$$

where ϕ is the filtering transform operator, such as the well-known wavelet, Γ is the soft or hard threshold operator [37]. For the optimization based methods, the denoising problem is usually formulated as follow:

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \frac{1}{2}||\boldsymbol{X} - \boldsymbol{Y}||^2 + \lambda\Phi(\boldsymbol{DX}). \tag{5}$$

Generally, the half quadratic splitting strategy [38] is adopted to optimize the problem (5). Thus, the original problem can be transformed into two easier sub-problems:

$$\begin{cases} \hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \frac{1}{2}||\boldsymbol{X} - \boldsymbol{Y}||^2 + \frac{\alpha}{2}||\boldsymbol{A} - \boldsymbol{DX} - \frac{\boldsymbol{J}}{\alpha}||^2 \\ \hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \frac{\alpha}{2}||\boldsymbol{A} - \boldsymbol{DX} - \frac{\boldsymbol{J}}{\alpha}||^2 + \lambda\Phi(\boldsymbol{A}), \end{cases} \tag{6}$$

in which the auxiliary variable $\boldsymbol{A}$ can be solved by the

$$\boldsymbol{A}^{(k+1)} = shrink_{\frac{\lambda}{\alpha}}(\boldsymbol{DX}^{(k)} + \frac{\boldsymbol{J}^{(k)}}{\alpha}), \tag{7}$$

where $\boldsymbol{D}$ is the sparse transformation operator, $\boldsymbol{J}$ can be regarded as the compensating variation, α is the regularization parameter, $shrink_{\frac{\lambda}{\alpha}}$ is the soft shrinkage operator, and k is the iteration number.

We can observe that Eq. (2), (4), (7) are very similar to each other. Their solutions all share the same format: a linear transformation and then non-linear activation function. This intrinsic similarity can partially explain why the deep model is also suitable for image restoration task.

Compared with the filter based methods, the CNN and optimization methods obtain the solution in a recursion man-

Table I
DETAIL ARCHITECTURE DESCRIPTION OF THE NOISE SUB-NETWORK. CR MEANS THE CONVOLUTIONAL + RELU. CB IS SHORT FOR CONVOLUTIONAL + BN. CBR DENOTES THE CONVOLUTIONAL + BN + RELU. THE NUMBER IS THE LAYER OF THE SUB-NETWORK.

Layer	Input	CR1	CBR2	CBR3	CB4	CBR5	CBR6	CB7	CBR8	CBR9	CB10	CBR11
Kernel Size	—	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3
Filter Number	—	64	64	64	64	64	64	64	64	64	64	64
Receptive Field	—	3	5	7	9	11	13	15	17	19	21	23
Image Size	40*40*1	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64

Layer	CB12	CBR13	CBR14	CBR15	CB16	CBR17	CBR18	CBR19	CB20	CBR21	CBR22	C23
Kernel Size	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3	3*3
Filter Number	64	64	64	64	64	64	64	64	64	64	64	1
Receptive Field	25	27	29	31	33	35	37	39	41	43	45	47
Image Size	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*64	40*40*1

Table II
QUANTITATIVE COMPARISON BETWEEN PLAIN-NET, ORACLE-NET AND PROPOSED NETWORK.

	Noise	Plain	Oracle	Proposed
PSNR	23.05	37.18	46.89	37.46
SSIM	0.2329	0.8921	0.9908	0.9001

ner. They gradually approximate to the desired solution with compensation ($\boldsymbol{P}^{(d)}$ and $\boldsymbol{J}^{(k)}$ can be understood as the compensation variables) in each iteration, which makes them more reasonable. The number of recursions depends on the depth of the deep model and the iterations of the optimization.

The conventional filter or optimization based methods use the hand-craft features, namely the ϕ and $\boldsymbol{D}$ are pre-defined. On the contrary, the transformation in CNN models are adaptively learned to implicitly fit the distribution of the training data, which makes them more professional for specific tasks.

3) From Practical Perspective: We visualize the extracted features by the filter method, optimization method, and the CNN. We chose the representative method: wavelet filter [7], total variational [9], and our CNN model for comparison. As shown in Fig. 2, the conventional methods could only extract the fixed pattern structures. The features in the CNN exhibit diversity from the singularity point, to various lines with different direction and thickness, to the rough profile.

D. The Architecture of Two-Stage CNN (TSCNN)

1) Advantageous of the Cascaded Sub-networks: As we have analyzed before, CNN is a more powerful tool for both the image and noise modeling. Directly end-to-end mapping the noisy image to the clean one may neglect the correlation between the image and noise. We name this network as the plain-Net [Fig. 3(a)]. Given the image decomposition framework, the image and noise mutually influence each other. The noises have specific distributions, which is worth taking into consideration along with the image content, and thus facilitate to decouple the image and noise components more thoroughly. A natural idea is to give the oracle/ground truth information about the noise to the image reconstruction network. And we name this network the oracle-Net [Fig. 3(b)]. However, the oracle/ground truth information about the noise is unknown in reality. In this work, we go further by learning two joint sub-networks for both image and noise component simultaneously, as shown in Fig. 3(c).

To verify the advantageous of the proposed network, we compare the three networks on the CT dataset with sigma = 20, as shown in Table II. We can observe that the amazing results by the oracle-Net are significantly higher than the others. The oracle-Net means that the inputs are the noise map and also the noisy image (the two components are concat together as the input, namely 40*40*2), and the output is the clean image. That is to say, the function of the network is to fit a subtractor (noisy image $\boldsymbol{Y}$ - noise map $\boldsymbol{N}$ = clean image $\boldsymbol{X}$). This phenomenon illustrates the **upper bound performance** represented by the oracle noise map. Without the noise map as guidance, namely the plain network, the quantitative performance is 37.18dB. We can see that with the estimated noise map as guidance, even it may be not exactly ground truth, it would also be beneficial to the denoising results. Here, we have 0.28dB improvement over the plain network. The proposed net and the oracle-Net import additional information about the noise for the network, thus they could better handle the noise. This motivates us to introduce the noise map estimation sub-network and design the cascaded CNN architecture.

In the first stage, we apply a CNN to learn an image to noise mapping, where the input is the noisy image and output is the estimated noise. The learned noise map which contains abundant information about the noise, such as the distribution (category of the noise) and intensity (noise level). In the second stage, we apply another CNN shared the same architecture as the first one to learn an image to image mapping. The estimated noise acts as a conditional map to indicate the image sub-network for better reconstruction. Thus, the proposed method is robust to different noise with various noise levels, which would be discussed in section III-D.

2) The Overall Architecture: In this work, we propose a two-stage CNN (TSCNN), as shown in Fig. 4. The network mainly contains three layers: convolutional, batch normalization (BN), and rectified linear unit (RELU). The convolutional layer is used to extract various features. The BN layer is incorporated for avoiding the gradient vanishing or divergence issue. And the RELU layer is utilized for pursuing sparsity and also for its highly nonlinear ability. No pooling layer or down-sampling operator is applied in our network, since this would inevitably cause the information lost for pixel level based image denoising task.

We use 3×3 filters throughout the whole net with stride 1. The filter number of each layer is fixed to 64. To avoid the boundary effect and preserve the spatial size, we pad

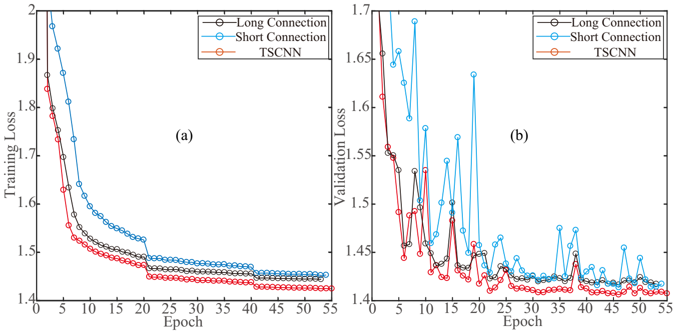


Figure 5. The effectiveness of the short-term and long-term connections. (a) Training loss. (b) Validation loss.

each layer with the same size as the original image. The receptive field of the network is much related to the depth of the network. Our training image size is 40×40. To fully utilize the whole contextual information of the given image, the depth of each sub-network is 23. Thus, the receptive field ($23 \times 2 + 1 = 47$) is slightly larger than the image size. To make it more clear, the architecture of the image/noise based sub-network is shown in Table I. It is worth noting that Table I shows the architecture of the noise sub-network, not the whole network. The noise sub-network and the image sub-network share a very similar architecture. Thus, we do not introduce the details of the image sub-network. The only difference between noise sub-network and the image sub-network is the first layer. The input of the noise sub-network is the noisy image 40*40*1, while the input of the image sub-network is both the noisy image and estimated noise map 40*40*2. And the details about the TSCNN are discussed in section III-D. It has been widely accepted that the deeper models involve more contextual information via a larger receptive field with better performance. However, when the depth of the model increases, the notorious gradient vanishing problem appears. In our network, we introduce both the short-term and long-term connections jointly to alleviate this issue.

3) Short-Term Connection: The gradient vanishing issue restricts us to train a very deep model with powerful representation. The essence of the gradient vanishing problem is that the gradient flow tends to be zero. The main reason is that the activation functions in each layer only respond to a certain percentage of the feature. The deeper the model is, the less the feature activates.

The residual block proposed by He et al. [32] is a powerful tool to accommodate this problem for two consecutive layers. The main idea of the residual learning, $\mathcal{F}(\boldsymbol{x}) = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{x}$, is to transform the original unreferenced mapping $\mathcal{F}(\boldsymbol{x})$ to residual mapping $\mathcal{H}(\boldsymbol{x})$. The residual blocks only need to learn the difference between its input and output. Such a simple reformulation by learning the sparse residual, not the image itself significantly facilitates to train a deep model, since the sparser gradient of the residual is easier to be propagated.

4) Long-Term Connection: Although the residual blocks facilitate the information to circulate via the short-term connections, the high-frequency details may still lose after very deep propagation. Moreover, it is worth noting that the image denoising task is a typical image to image transformation task, in which most of the information should be preserved and only the sparse noise component needs to be removed.

This motivates us to introduce the long-term skip connections to promote the information propagation. Skip connections [39] between layers have been long studied in neural networks, to improve the flow of information. In our task, on the one hand, the long term connection works as a memory module to preserve the similarity between the input and output, which is similar to that of the fidelity term in the optimization model. On the other hand, it also can compensate for the information loss during the propagation and enhance high-frequency details.

To validate the effectiveness of the long and short term connections, in Fig. 5, we show the training and validation loss curve with only long-term connection (black curve), with only short-term connection (blue curve), and with both of them (red curve). Compared the red curve with black curve and blue curve, we can conclude that both the long-term connection and short-term connection benefit to the final training procedure. Moreover, in low-level image-to-image translation task, the long-term connection is more important than short-term connection.

5) Fine-tuning Strategy: In our work, we first trained our model on the Berkeley Segmentation dataset (BSD)[1] with 204,800 sub-samples with the size 40×40. This dataset contains abundant image structures such as the fine textures and large scale edges with different scale and direction, which has been widely used as a fair benchmark for training [19], [40], [41]. Then, we fine-tuned the trained model on particular image datasets with 90 percent for training and the remaining 10 percentage for testing. The reasons are three-fold. On the one hand, we have not enough medical training samples, such as the clean and noisy ultrasound image pairs. On the other hand, natural images contain abundant multiscale information. That is to say, the trained model on the natural image could be easily transferred to arbitrary images. Last but not least, the noise sub-network mainly learns the representation of the noises, which is applicable to both the natural and medical images.

E. Training Details

The training of the network is to minimize the Eq. 3 and learn the parameters $\{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, ..., \boldsymbol{W}^{(D)}\}$. We introduce the ADAM [36] solver to optimize the problem. For the initialization of the parameters, we follow the Xavier method proposed by [42]. The learning rate is initialized as 0.0005 and decay 1/2 every 20 epochs. The momentum is 0.9 and a mini-batch size is 128. And the training criterion is consistent for all kinds of noises. We use MatConvnet [43] to implement our network. In our work, we jointly train the cascaded network in an end-to-end manner which is similar to the multitask learning. On the one hand, the two tasks are tightly coupled, which is beneficial to each other. On the other hand, the individual training of each sub-network and fine-tuning the whole cascaded network would be time-consuming.

[1] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

Table III
A COMPARISON OF STATE-OF-THE-ART MEDICAL IMAGE DENOISING METHODS AND THEIR PROPERTIES.

Method	Task	Scalability	Vectorization	Information	Prior	Modeling	Speed	Performance
OWT	Denoising	No	No	Local	Hand-crafted	Image	★★★★	★★★
BM3D	Denoising	No	No	Local+Nonlocal	Hand-crafted	Image	★★★⯨	★★★★
WNNM	Denoising	No	Yes	Local+Nonlocal	Hand-crafted	Image	★★	★★★★
DnCNN	Denoising	Yes	No	Local+Global	Learned	Noise	★★★★★	★★★★⯨
OBNLM	Despeckling	No	No	Local	Hand-crafted	Image	★★⯨	★★★
NLLR	Despeckling	No	Yes	Local+Nonlocal	Hand-crafted	Image	★	★★★⯨
IDCNN	Despeckling	Yes	No	Local+Global	Learned	Noise	★★★★★	★★★★⯨
WFFT	Destriping	No	No	Local	Hand-crafted	Image	★★★★	★★★⯨
VSNR	Destriping	No	No	Local	Hand-crafted	Noise	★★★	★★★★⯨
DLS-NUC	Destripnig	Yes	No	Local+Global	Learned	Noise	★★★★⯨	★★★★
MMF	Deringing	No	No	Local	Hand-crafted	Image	★★★⯨	★★★
VSC	Deringing	No	No	Local	Hand-crafted	Image	★★⯨	★★★
TSCNN	Comprehensive	Yes	No	Local+Global	Learned	Image+Noise	★★★★⯨	★★★★★

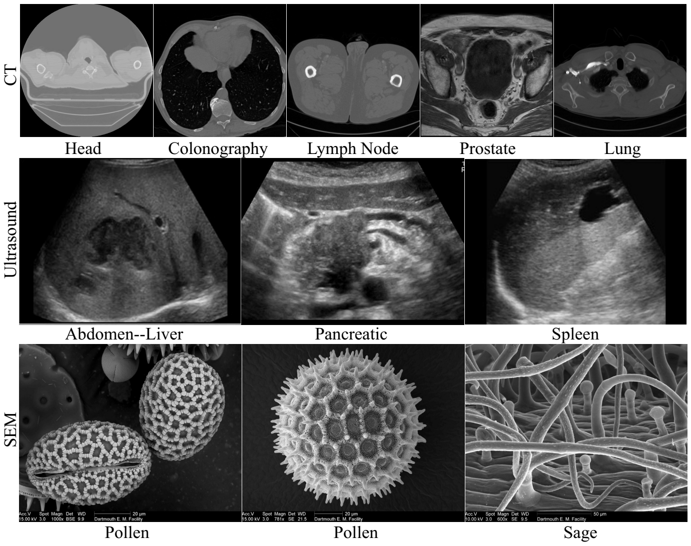


Figure 6. Illustration of the training and testing samples. The first row shows the representative CT images with five different parts of human. The second row shows the representative ultrasound images with three different parts of human. The third row shows the representative SEM images with two different kinds of plants.

III. Experimental Results and Discussion

A. Experimental Setting

1) Datasets: We use three kinds of medical image datasets for different noises: the random Gaussian noise in CT, the random speckle noise in the ultrasound image, the line pattern noise in SEM, and the ring pattern noise in CT. The CT dataset is downloaded from National Biomedical Imaging Archive (NBIA)[2], typically patients related by a common disease (such as lung, Lymph, and Prostate cancer) with various image modalities (such as CT and MRI). DICOM is the primary file format used by NBIA for image storage. We collected 1000 CT images with the size 256×256. The clinical ultrasound dataset[3] comes from the teaching files from the Gelderse Vallei Hospital in Ede, the Netherlands, which included a large number of general ultrasound cases collected over the years by the radiologists and ultrasound technicians of the hospital. The SEM dataset is downloaded from Dartmouth College Electron Microscope Facility[4], which included various species such as the blood cells and pollen. We collected 300 images and resized them to size 512×640.

Table IV
QUANTITATIVE RESULTS OF DIFFERENT METHODS UNDER DIFFERENT GAUSSIAN NOISE LEVELS IN CT.

Noise Level	Index	Methods						
		Noisy	OWT	BM3D	WNNM	DnCNN	TSCNN-NFT	TSCNN
10	PSNR	28.94	38.22	41.13	41.16	40.59	40.68	**41.85**
	SSIM	0.4924	0.9184	0.9477	0.9502	0.9453	0.9464	**0.9556**
20	PSNR	23.05	34.79	37.84	37.97	37.01	37.46	**39.16**
	SSIM	0.2329	0.8659	0.9175	0.9212	0.8719	0.9108	**0.9338**
30	PSNR	19.60	32.83	35.76	36.16	34.88	35.37	**37.18**
	SSIM	0.1363	0.8280	0.8896	0.9004	0.8176	0.8546	**0.9181**
50	PSNR	15.39	30.47	33.16	33.74	32.57	32.85	**34.72**
	SSIM	0.0655	0.7759	0.8537	0.8729	0.7699	0.7831	**0.8887**

Table V
QUANTITATIVE RESULTS OF DIFFERENT METHODS UNDER DIFFERENT SPECKLE NOISE LEVELS IN ULTRASOUND.

Noise Level	Index	Methods				
		Noisy	OBNLM	NLLR	IDCNN	TSCNN
0.01	PSNR	23.37	28.57	29.55	34.25	**34.49**
	SSIM	0.5747	0.8578	0.7983	0.9272	**0.9330**
0.02	PSNR	22.56	27.71	28.01	32.71	**33.04**
	SSIM	0.5201	0.8335	0.7477	0.9086	**0.9144**
0.04	PSNR	19.66	26.41	26.78	31.16	**31.57**
	SSIM	0.4150	0.7396	0.7129	0.8843	**0.8938**

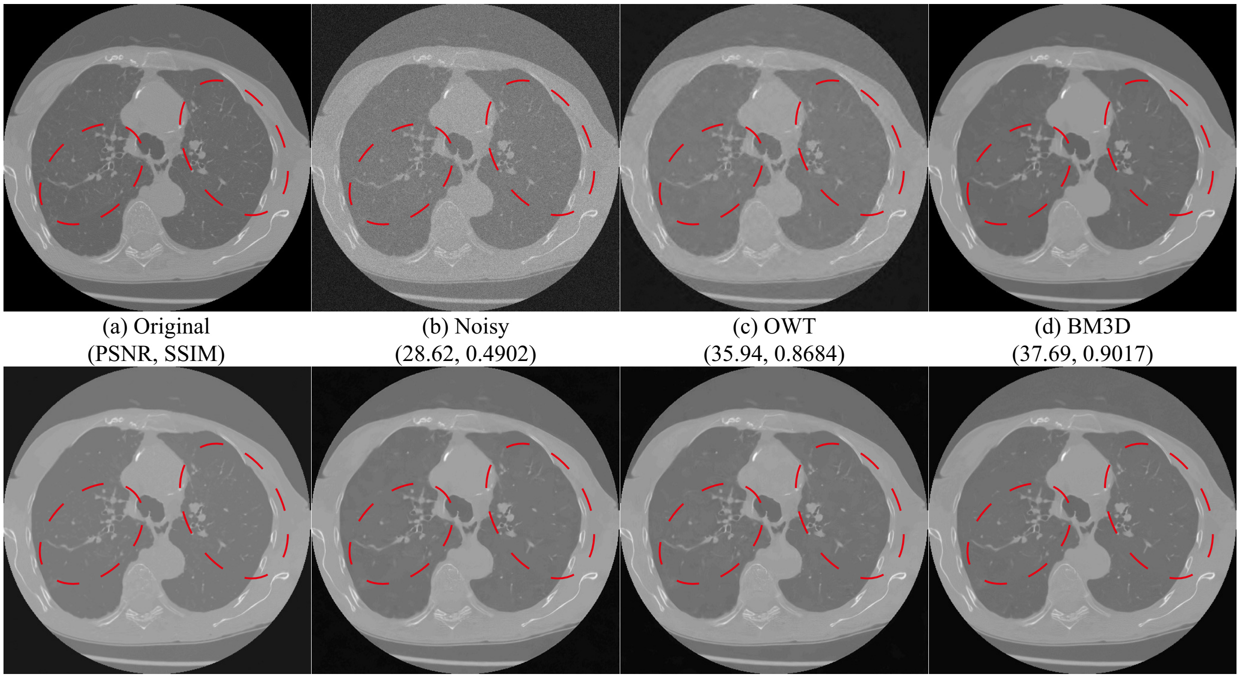
We show some of the representative medical images we collected in Fig. 6. The first row shows the representative CT images with five different parts of a human. The second row shows the representative ultrasound images with three different parts of a human. The third row shows the representative SEM images with two different kinds of plants. For example, we collect the CT images with different human parts, such as Head, Colonography, Lymph Node, Prostate, Lung. We can see that these CT images exhibit different structural shape with different background.

2) Compared Methods: We compare the proposed method with state-of-the-art denoising methods for different noises. For the Gaussian noise, the comparison methods include the filter based OWT [44] and BM3D [15], the low-rank based

[2] https://public.cancerimagingarchive.net/ncia/searchMain.jsf

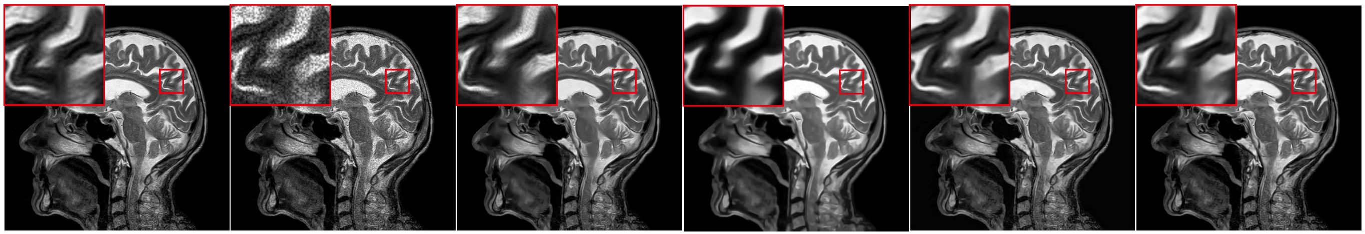
[3] http://www.ultrasoundcases.info

[4] http://www.dartmouth.edu/~emlab/gallery/



(a) Original (PSNR, SSIM)
(b) Noisy (28.62, 0.4902)
(c) OWT (35.94, 0.8684)
(d) BM3D (37.69, 0.9017)
(e) WNNM (37.82, 0.9071)
(f) DnCNN (37.28, 0.8796)
(g) TSCNN_NOfinetune (37.43, 0.8989)
(h) TSCNN_finetune (38.22, 0.9125)

Figure 7. Comparison for simulated Gaussian noise. (a) Clean CT image. (b) Gaussian image. Denoising results by (c) OWT, (d) BM3D, (e) WNNM, (f) DnCNN, (g) TSCNN without finetuning on CT, (h) TSCNN with finetuing on CT.



(a) Original (PSNR, SSIM)
(b) Noisy (24.52, 0.7811)
(c) OBNLM (31.68, 0.9090)
(d) NLLR (29.04, 0.8322)
(e) IDCNN (34.20, 0.9449)
(f) TSCNN (34.45, 0.9477)

Figure 8. Comparison for simulated speckle noise. (a) Clean image. (b) Speckle image. Despeckling results by (c) OBNLM, (d) NLLR, (e) IDCNN, (f) TSCNN.

WNNM [45], and the deep learning based DnCNN [19]. For the speckle noise, we compare the TSCNN with filter based OBNLM [46], low-rank based NLLR [47], and deep learning based IDCNN [48] methods. For the stripe noise, we compare with the filter based WFFT [49], variational based VSNR [50], and deep learning based DLS-NUC [51] methods. For the ring noise, we introduce the filter based MMF [52], variational based VSC [8] and deep learning based DLS-NUC [51] methods. The quantitative assessments PSNR and SSIM [53] are introduced to give an overall evaluation. We have obtained the source code of the compared methods from the homepage of the authors, and set the parameters following the rules in the compared papers with default parameters to obtain the best results. For the reproduction of our research, the training code will be available on our homepage[5].

Here we give a brief introduction to each of them. We give

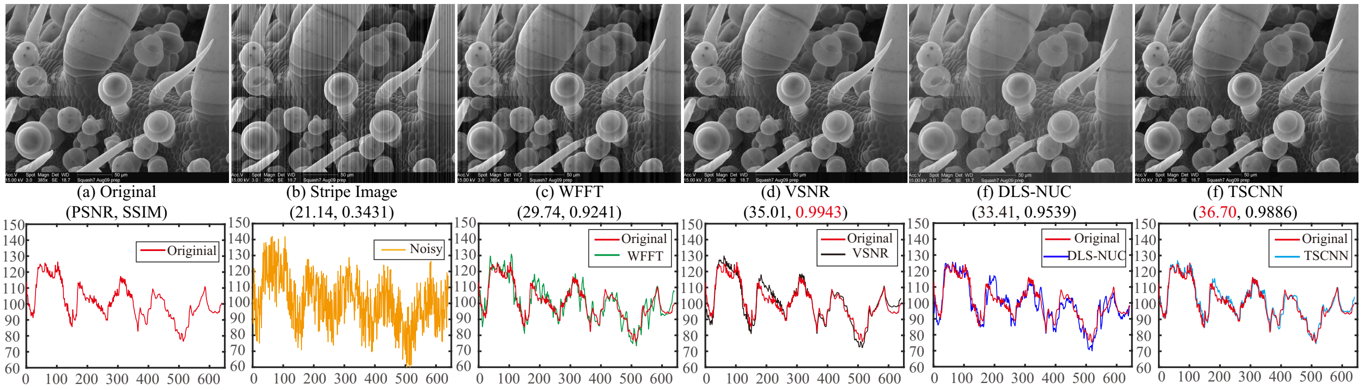
[5] http://www.escience.cn/people/changyi/index.html

Table VI
QUANTITATIVE RESULTS OF DIFFERENT METHODS UNDER DIFFERENT STRIPE NOISE LEVELS IN SEM.

Noise Level	Index	Methods				
		Noisy	WFFT	VSNR	DLS-NUC	TSCNN
10	PSNR	33.25	40.40	41.61	38.59	**42.90**
	SSIM	0.8537	0.9841	0.9909	0.9771	**0.9942**
20	PSNR	27.27	35.32	39.38	36.77	**39.61**
	SSIM	0.6633	0.9645	0.9823	0.9703	**0.9895**
40	PSNR	21.31	29.80	35.24	33.75	**35.75**
	SSIM	0.4294	0.9191	**0.9799**	0.9482	0.9568

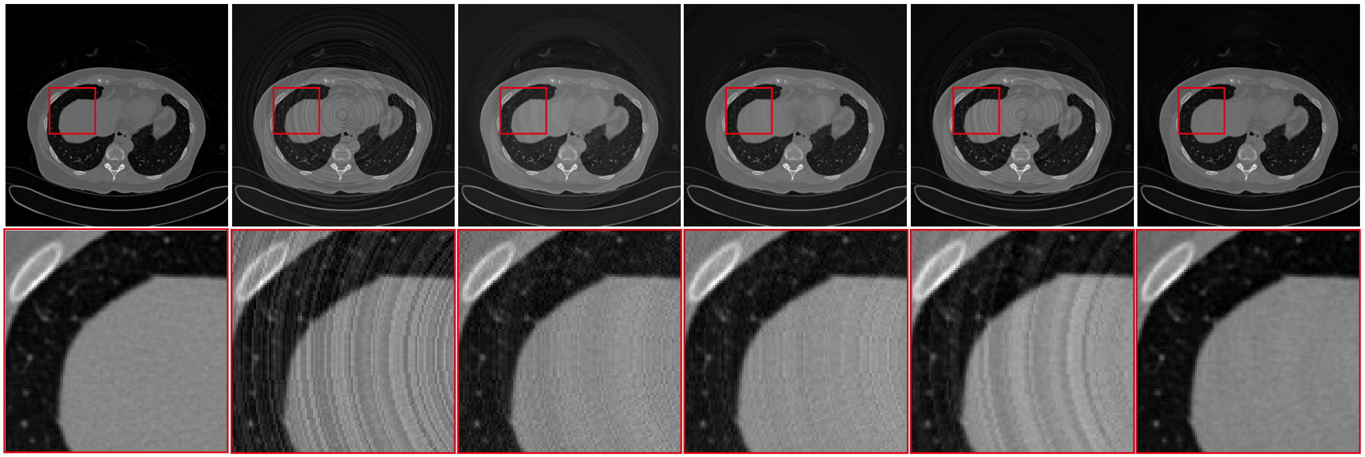
Table VII
QUANTITATIVE RESULTS OF DIFFERENT METHODS UNDER DIFFERENT RING NOISE LEVELS IN CT.

Noise Level	Index	Methods				
		Noisy	MMF	VSC	DLS-NUC	TSCNN
10	PSNR	34.44	38.48	38.58	38.69	**45.99**
	SSIM	0.8235	0.9344	0.9429	0.9496	**0.9864**
20	PSNR	28.46	33.77	33.89	33.94	**43.11**
	SSIM	0.5965	0.8226	0.8388	0.8675	**0.9791**



(a) Original (PSNR, SSIM)
(b) Stripe Image (21.14, 0.3431)
(c) WFFT (29.74, 0.9241)
(d) VSNR (35.01, 0.9943)
(f) DLS-NUC (33.41, 0.9539)
(f) TSCNN (36.70, 0.9886)

Figure 9. Comparison for simulated stripe noise level $\{-40, 40\}$. (a) Clean SEM image. (b) Stripe image. Destriping results by (c) WFFT, (d) VSNR, (e) DLS-NUC, (f) TSCNN. The horizontal axis denotes the line number, and the vertical axis means the corresponding mean value.



(a) Original (PSNR, SSIM)
(b) Noisy (29.53, 0.6277)
(c) MMF (33.03, 0.7937)
(d) VSC (36.64, 0.8754)
(e) DLS-NUC (34.52, 0.8684)
(f) TSCNN (40.92, 0.9521)

Figure 10. Comparison for simulated ring noise level $\{-20, 20\}$. (a) Clean CT image. (b) Ring image. Deringing results by (c) MMF, (d) VSC, (e) DLS-NUC, (f) TSCNN.

a comprehensive properties comparison in Table III. The task means what kinds of noise do they process in the work. The scalability denotes whether they can be extended to all other noise removal tasks. The vectorization is whether the input 2-D image/patch has been transformed into 1D vectors or not (Preserving the structure information intact is very important in image restoration). The information represents that what kinds of information have been utilized (The local based methods mainly utilize the information in a local patch/neighborhood. The nonlocal based methods take the nonlocal self-similarity into consideration. The global based methods take advantage of the whole contextual information). The prior means whether they model the different dataset adaptively or fixedly. The learned prior makes the method quite flexible for different images, while the hand-crafted prior may lose its professional ability for the specific image. We can observe that most of the existing methods are still based on hand-crafted prior. We hope that the learning based method could receive more attention. The modeling can be mainly classified into two categories: the images or the noises. That is to say, some methods mainly estimate the clear image, while some other methods estimate the noise component instead. It is worth noting that different from all competing methods, our TSCNN models the image and the noise component simultaneously from an image decomposition perspective. The speed and the performance represent the running time and the denoising results of each method, respectively.

B. Simulated Results

Figure 7 to 10 show the comparison results of our TSCNN with the state-of-the-art methods for different noise cases. It can be observed that our TSCNN obtain the best performance in terms of the detail preserving, noise removal, visual appearance and also the quantitative indexes (marked by the red). Four points are worthy to be noticed. First, our method is quite flexible for various kinds of noises with different distributions, including both the random and structural noise, while previous methods only work well for specific noise. Second, the fine-tuning strategy of our pre-trained model (on BSD natural image) to specific medical image greatly boosts the final restoration performance, as shown in Fig. 7(g) and (h), which also significantly reduce the training samples of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2019.2925881, IEEE Transactions on Instrumentation and Measurement

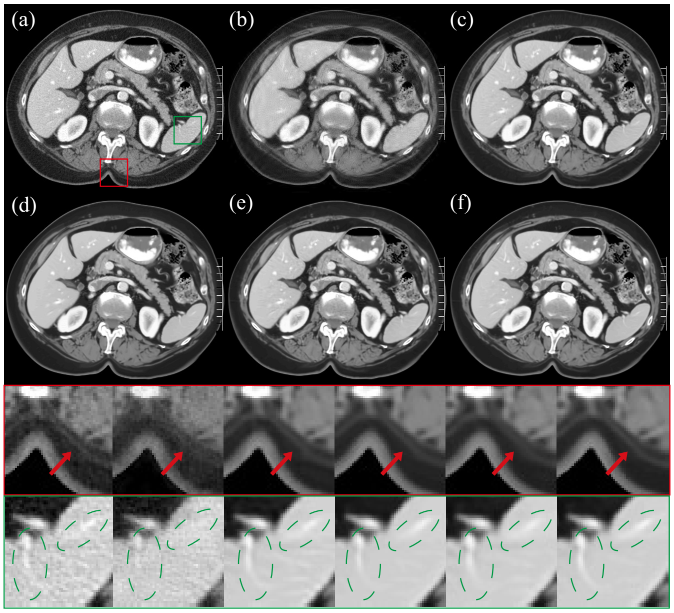


Figure 11. Comparison for real Gaussian noise. (a) Gaussian image. Denoising results by (b) OWT, (c) BM3D, (d) WNNM, (e) DnCNN, (f) TSCNN.

Table VIII
QUANTITATIVE COMPARISON BETWEEN TSCNN AND DLS-NUC FOR DIFFERENT IMAGE CATEGORIES UNDER DIFFERENT NOISE LEVELS.

Data	Noise Level	10		20		30		40	
	Index	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set12	DLS-NUC	34.50	0.9679	32.98	0.9609	31.81	0.9541	30.69	0.9423
	TSCNN	**41.82**	**0.9935**	**38.87**	**0.9894**	**36.86**	**0.9865**	**35.30**	**0.9839**
NIR	DLS-NUC	39.67	0.9798	37.49	0.9734	35.71	0.9677	33.96	0.9550
	TSCNN	**43.22**	**0.9928**	**40.25**	**0.9890**	**37.91**	**0.9861**	**37.09**	**0.9832**

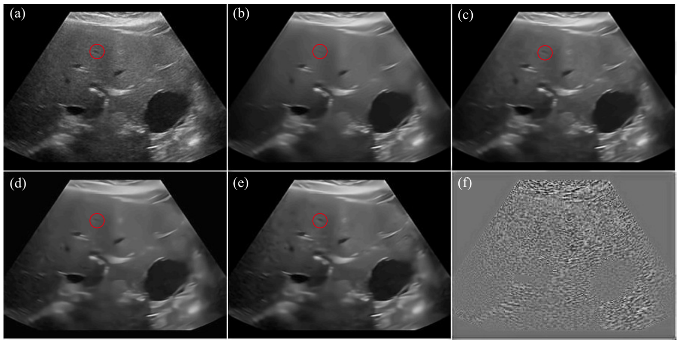


Figure 12. Comparison for real speckle noise. (a) Speckle image. Despeckling results by (b) OBNLM, (c) NLLR, (d) IDCNN, (e) TSCNN. (f) Estimated speckle by TSCNN.

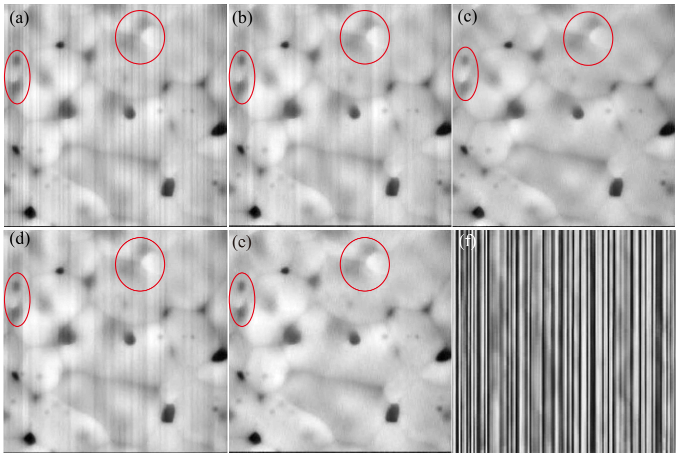


Figure 13. Comparison for real stripe noise. (a) Stripe image. Destriping results by (b) WFFT, (c) VSNR, (d) DLS-NUC, (e) TSCNN. (f) Estimated stripe by TSCNN.

medical images. Third, from the region of interests marked by the red ellipse and square in Fig. 7, 8, 10, we can observe that the noises have been consistently suppressed, and different scale structure information is well preserved. Last, in Fig. 9, in terms of both the 2D image and 1D mean-cross profile [54], TSCNN achieves a better destriping result than the competing methods.

We report our quantitative assessments accordingly from Table IV to VII. The highest PSNR and SSIM values are highlighted in bold. We have the following observations. First, the proposed method achieves the highest PSNR and SSIM values in most cases, which verify the effectiveness of the CNN for various noises modeling in medical images. Second, the difficulty of removing these noises for our method gradually arises from the ring, Gaussian, stripe, to speckle noise. Third, the TSCNN significantly improves the records in terms of the speckle and ring noise removal, and outperforms the state-of-the-art despeckling and deringing methods with a large marginal.

It is worth noting that, we re-train the DnCNN and IDCNN on our datasets, while the DLS-NUC does not provide the source training code but with the trained model on infrared image. According to our experiments for CT images Gaussian noise removal, as shown in Table IV and Fig. 7, the improvement of the fine-tuning on the specific image would boost improvement about PSNR $1 \sim 2dB$ and SSIM $1\% \sim 10\%$ for different noise levels. In Table VI and VII, we can find that the proposed TSCNN outperforms the DLS-NUC with approximate 3dB and 7dB, respectively. That is to say, the proposed TSCNN could still work better than the DLS-NUC even with the fine-tuning. We further give an overall comparison on near infrared image NIR [55] and natural image Set12 [19] datasets. For NIR dataset, we randomly choose 40 NIR with size 1024*680 as our testing dataset. For natural image, there are 12 common images with size 256*256 or 512*512 as our testing dataset. The comparison results are listed in Table VIII. We can observe that the proposed method consistently outperforms the DLS-NUC on different image categories, which strongly support the superiority of the TSCNN over the DLS-NUC. It is worth noting that even our model has never 'seen' the infrared image before, namely no training or fine-tuning on the infrared image, the proposed TSCNN still works satisfactorily in terms of quantitative and qualitative indexes. This phenomenon reveals that the proposed model has captured the intrinsic line pattern feature no matter where the stripe noise exists.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2019.2925881, IEEE Transactions on Instrumentation and Measurement

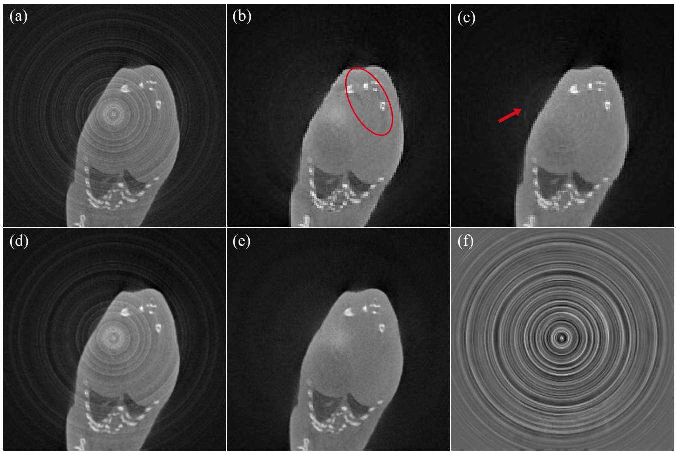
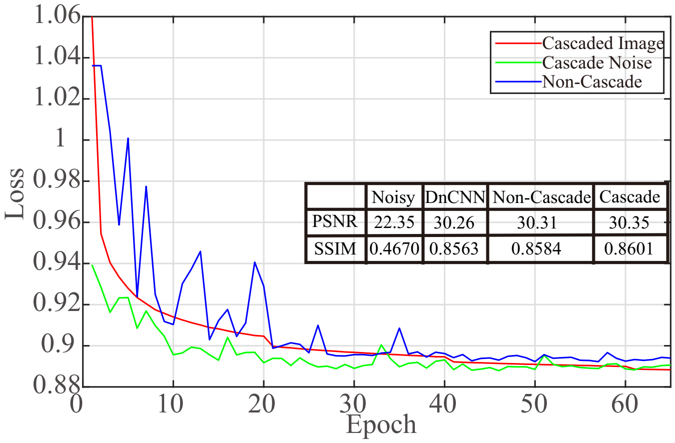


Figure 14. Comparison for real ring noise. (a) Ring image. Deringing results by (b) MMF, (c) VSC, (d) DLS-NUC, (e) TSCNN. (f) Estimated ring by TSCNN.



	Noisy	DnCNN	Non-Cascade	Cascade
PSNR	22.35	30.26	30.31	30.35
SSIM	0.4670	0.8563	0.8584	0.8601

Figure 15. The effectiveness of cascaded architecture. We compare the training loss and quantitative performance with/without the cascaded architecture.

C. Real Medical Image Results

Figure 11 presents a real Abdominal/Pelvic CT image via a thin beam of X-ray from the St. Elizabeth's Medical Center[6]. Figure 12 shows a real clinical ultrasound image with multiple liver cysts obtained from [47]. Figure 13 presents a real SEM imaging on a sintered specimen of CeO_2 [56]. Figure 14 shows a real CT image of a rat reconstructed by FDK algorithm [8]. It can be observed that the results of the proposed method exhibit good visual quality with fewer artifacts than the results obtained with the other methods. The noises are significantly suppressed and the detail information is perfectly preserved by the proposed method, while the existing methods either smooth the image details (such as Fig. 12(b) and Fig. 13(c)) or contain the residual noise (such as Fig. 13(b)). These experiments demonstrate the effectiveness of our method for real noisy images.

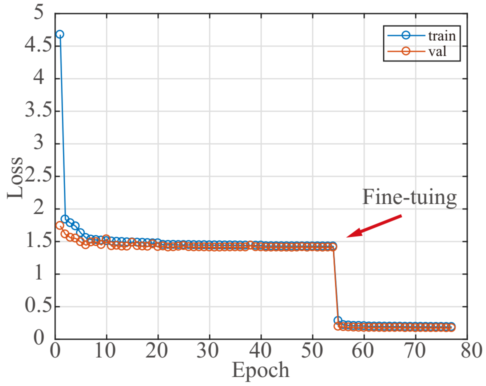


Figure 16. The effectiveness of the fine-tuning. After the fine-tuning, the training loss drops drastically.

D. Discussion

1) Effectiveness of the Cascaded Architecture: In our cascaded CNN, the first noise based sub-network works as a noise estimator which tells the second image sub-network about the noise level and category information of the noise. With sufficient knowledge known in advance via the first sub-network, the second sub-network could remove different kinds of noises categories and noise levels. On the contrary, the previous methods may fail to achieve this goal. Here, we demonstrate this property of our work from two aspects: training loss and noise estimation.

In Fig. 15, we show the training loss of the non-cascaded model (namely only the noise-CNN sub-network, the blue curve) and the proposed cascaded model (green curve for the loss of noise-based sub-network and red curve for the loss of image-based sub-network). We can observe that the cascaded model could further boost the training procedure with lower loss than that of the non-cascaded model. Moreover, we show the quantitative results of both cascade and non-cascade model, and we take the DnCNN [19] as a baseline. We test them on BSD68 with Gaussian noise $\sigma = 20$. From the quantitative results, we can also conclude that our cascaded model could benefit the training thus with better denoising performance.

2) Effectiveness of the Fine-tuning: The fine-tuning strategy (training on a larger dataset first and then fine-tune on the smaller target dataset) has been widely used for various vision tasks. As far as we know most of the detection methods heavily rely on the pre-trained backbone network on Imagenet. In Fig. 16, we show the training and validation loss before and after fine-tuning. We observe that after 54 epochs both the training and validation loss on the BSD decreased slowly. Here, we fine-tune the pre-trained model on CT dataset. The loss dropped suddenly and gradually converged in a few numbers of epochs. We also compare their visual results in Fig. 7 (g) and (h). These results strongly prove the effectiveness of the fine-tuning strategy.

3) Robustness to Different Noise/Image Categories: Explicitly modeling the distributions of the complex noises in medical images, such as the mixed noise categories and noise levels, is extremely hard for previous methods. We show our

[6] https://www.semc.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2019.2925881, IEEE Transactions on Instrumentation and Measurement

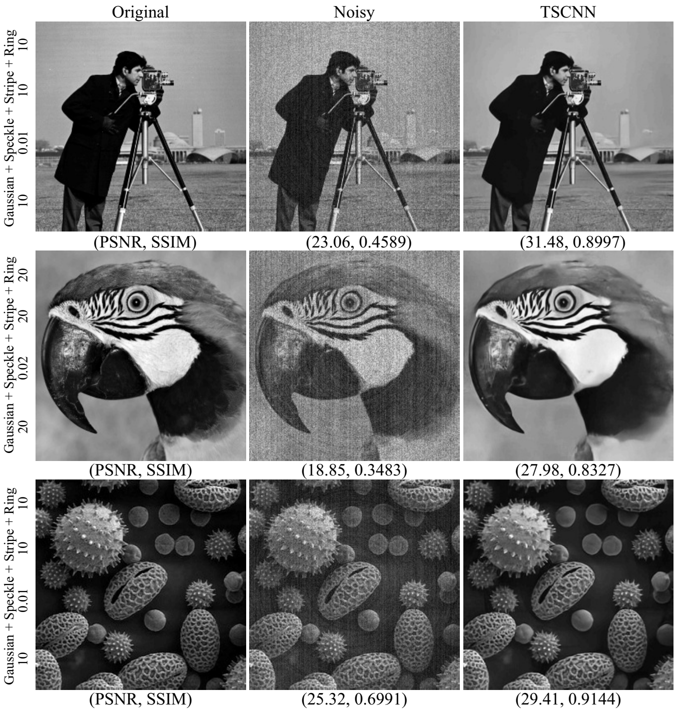


Figure 17. Effectiveness of the proposed method with single model for different noise levels, noise types and image types.

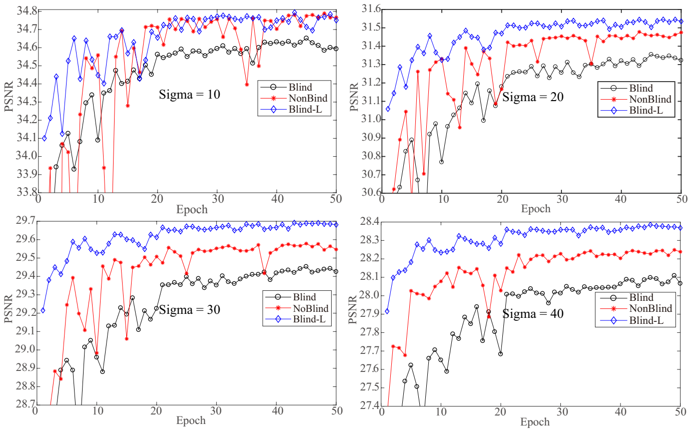
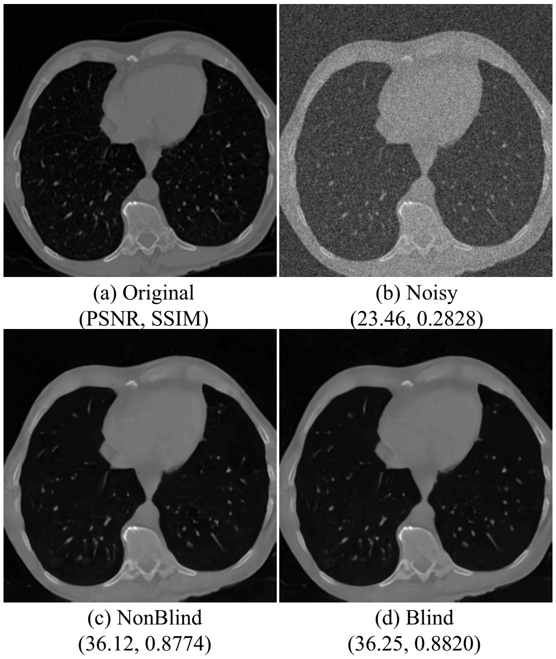


Figure 18. The comparison between the specific and general models for different noise levels. The black, red and green curve denote the trained model for single noise level (nonblind), mixed noise levels (blind) and mixed noise levels with larger training dataset (blind-L), respectively.


(a) Original
(PSNR, SSIM)

(b) Noisy
(23.46, 0.2828)

(c) NonBlind
(36.12, 0.8774)

(d) Blind
(36.25, 0.8820)

Figure 19. The visual comparison between the specific and general model. (a) Original CT image. (b) The image with Gaussian noise Sigma = 20. Denoising results by (c) the specific model and the general model.

method is robust to any mixed noise with more complex noise distribution. In Fig. 17, we perform an experiment to train **one single model** for mixed noise categories (four kinds of common noise: Gaussian noise + Speckle noise + Stripe noise + Ring noise) and mixed noise levels (Gaussian noise level 10/20, Speckle noise level 0.01/0.02, Stripe noise level 10/20, Ring noise level 10/20). We can clearly observe that our method consistently obtains the visual pleasure results with satisfactory quantitative index, which strongly validates the effectiveness and robustness of TSCNN to any mixed noise and image type. As far as we know, there are few methods considering such realistic but challenging noises from a single image. We do believe that our method could be also well applied to other artifacts in medical images, such as the metal artifacts in CT.

4) Robustness to Different Noise Levels: The noise model not only refers to a statistical probability distribution, but also its noise level. It is worth noting that, for the conventional filtering/optimization based method, the noise level is usually known in advance. They control the denoising strength by manually adjusting the regularization parameter, which is associated with the noise level. For most of the previous methods, not only the noise model need to be known, but also the noise level has to be known in advance. If not, their performance will decrease dramatically.

On the contrary, our deep cascaded model is not only robust to the noise distribution but also the noise level. We train one single model for different noise levels. We show the comparison results of the general model with the specific noise level model. We can observe the proposed model trained for all noise levels still works well. Compared with the separate model (red curve), the results of blind noise model (black curve) degenerate $0.1 \sim 0.2dB$ in all noise levels, as shown in Fig. 18. However, when we enlarge the training datasets with four times, the results of blind noise model (green curve) could even increase $0.1 \sim 0.2dB$ in all noise levels. This interesting phenomenon demonstrates two things. On the one hand, the proposed model is very robust to different noise levels. On the other hand, the limited dataset has not utilized the full potential of the proposed neural network. Our results could be significantly improved with the larger training dataset. We also give a visual comparison between the blind and nonblind model, as shown in Fig. 19. Both the visual appearance and the quantitative results have slight improvement. The first sub-

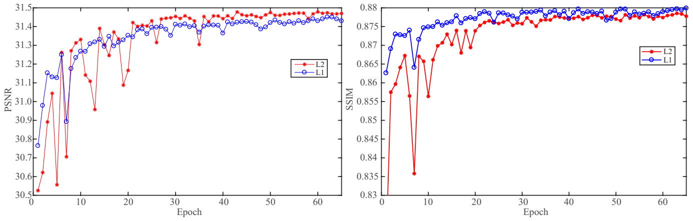


Figure 20. The comparison between the L_2 and L_1 loss. The results are very close to each other.

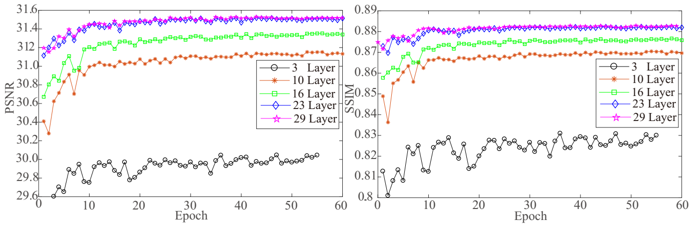


Figure 21. The influence of the depth of the proposed network. Left is the PSNR vs the epoch, and right is the SSIM vs the epoch.

network can be regarded as a noise estimator with relatively high accuracy, so that the image estimation sub-network know the noise level in advance.

5) The Choice of the Loss Functions: Here, we compare the difference between the L_2 and L_1 loss, as shown in Fig. 20. We can observe that both the PSNR and SSIM value of the L_2 and L_1 are very close to each other. That is to say, the reconstruction loss functions only influence the denoising result a little. Thus, we employ the L_2 loss in Eq. (3) for both the image and noise reconstruction.

6) The Influence of the Network Depth: We explore the influence of the network depth for the final performance. In Fig. 21, we compare the PSNR and SSIM value of the model with different depths. Here, we choose the model with 3, 10, 16, and 23 layers of each sub-network as a representation. The 3, 10, 16, and 23 layers model correspond to the proposed model with 0, 1, 2, and 3 long-term connections, respectively (Please refer to the Fig. 4.). For example, for the 3 layer model, it means that we cut off all the intermediated layers in 23 layers sub-network (only the first and last two layers are left). It can be clearly observed that the deeper the model is, the better the denoising results are. Then, we increase the layer of each sub-network to 29 layers (purplish red curve). Compared the network of 23 layers with 29 layers, we can hardly observe the increase of the PSNR and SSIM value in Fig. 21. The training time of each depth (3, 10, 16, 23, and 29) is 0.2, 0.8, 1.4, 2.2, and 2.8 days, respectively. The memory space of each depth model (3, 10, 16, 23, and 29) is 0.3M, 5.5M, 10.6M, 16.9M, and 20.1M. Therefore, we choose 23 layers to obtain a satisfactory balance between performance and resource consumption.

7) Testing Speed: In Table IX, we show the running test time of all competing methods. To give a fair comparison, we uniformly resize different kinds of medical images into 512×512. We perform the experiment on MATLAB 2017a, an Intel i7 CPU at 3.6 GHz, an NVIDIA 1080Ti GPU, and 32-GB memory. We can see that the non-local based methods are the slowest mainly caused by the non-local patch searching, such as the WNNM and NLLR. Next is the variational based methods (VSNR and VSC), due to the iterations. The filtering based methods are comparatively fast in the FFT domain (OWT and WFFT). Normally, the CNN based methods obtain the best performance in running times due to their simple forward and parallel computation in GPU. As for the training time, the pre-training time of our model is normally within two days and fine-tuning for half a day.

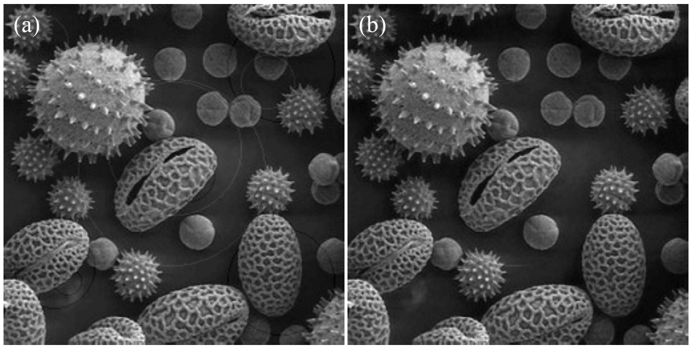


Figure 22. Ring removal with arbitrary position. (a) Ring image. (b) TSCNN.

Table IX
RUNNING TIME (SEC) OF THE COMPARING METHODS ON 512*512 IMAGE.

Method	OWT	BM3D	WNNM	OBNLM	DnCNN(CPU/GPU)
Time	0.23	2.07	48.72	11.94	4.32/0.003
NLLR	WFFT	VSNR	MMF	VSC	TSCNN(CPU/GPU)
1259.41	0.24	8.93	2.85	16.57	15.56/0.025

8) Interesting Extension: In Fig. 22, we show the result of the TSCNN model trained with ring circle in the middle. We can observe that the test ring circle with arbitrary *unknown* positions can also be well removed by our method. The previous methods all need to know the center of the circles. Such an interesting phenomenon demonstrates the CNN has learned to capture the features of the circle pattern in the image automatically. This may provide a new insight for the vesicle structures detection and removal in transmission electron microscope images [57].

IV. Conclusion

In this paper, we propose to remove the noises in medical images from the image decomposition perspective. Different from previous works, the noise component and image component are treated equally in our work, and a two-stage convolutional neural network is proposed to model both the image and noise simultaneously. Instead of explicitly modeling the complex distribution of various noises and multi-modality medical images, the deep model could automatically figure out the distribution of the specific noise and image from a data-driven viewpoint. The proposed cascaded CNN model benefits us to handle different noise categories and noise levels adaptively. To facilitate the training, we introduce both the short-term and long-term connections in the network for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2019.2925881, IEEE Transactions on Instrumentation and Measurement

better information propagation. Moreover, we apply the fine-tune strategy to alleviate the lack of medical images issue. Extensive simulated and real medical image datasets have been tested. Experimental results demonstrate that the proposed method is very effective for various noises, and outperforms the state-of-the-art methods.

Our work shows that CNN is a powerful tool for modeling the noises and multi-modality images with fast test speed. We believe other low-level image processing problems such as the deblurring and super-resolution tasks could also benefit from the deep model. Moreover, it is interesting to see more advanced deep models for medical image analysis. For example, the 3D CNN could well handle multi-slice data with temporal information.

References

[1] M. Karaman, M. A. Kutay, and G. Bozdagi, "An adaptive speckle suppression filter for medical ultrasonic imaging," *IEEE Trans. Med. Imag.*, vol. 14, no. 2, pp. 283–292, 1995.

[2] T. Loupas, W. McDicken, and P. L. Allan, "An adaptive weighted median filter for speckle suppression in medical ultrasonic images," *IEEE Trans. Circuits and Syst.*, vol. 36, no. 1, pp. 129–135, 1989.

[3] F. Attivissimo, G. Cavone, A. M. L. Lanzolla, and M. Spadavecchia, "A technique to improve the image quality in computer tomography," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 5, pp. 1251–1257, 2010.

[4] M. Georgiev, R. Bregović, and A. Gotchev, "Fixed-pattern noise modeling and removal in time-of-flight sensing," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 4, pp. 808–820, 2016.

[5] F. Russo, "An image-enhancement system based on noise estimation," *IEEE Trans. Instrum. Meas.*, vol. 56, no. 4, pp. 1435–1442, 2007.

[6] A. Mencattini, M. Salmeri, R. Lojacono, M. Frigerio, and F. Caselli, "Mammographic images enhancement and denoising for breast cancer detection using dyadic wavelet processing," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 7, pp. 1422–1430, 2008.

[7] I. Firoiu, C. Nafornita, J.-M. Boucher, and A. Isar, "Image denoising using a new implementation of the hyperanalytic wavelet transform," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 8, pp. 2410–2416, 2009.

[8] L. Yan, T. Wu, S. Zhong, and Q. Zhang, "A variation-based ring artifact correction method with sparse constraint for flat-detector ct," *Phys. Med. Biol.*, vol. 61, no. 3, p. 1278, 2016.

[9] W. Zhao and H. Lu, "Medical image fusion and denoising with alternating sequential filter and adaptive fractional order total variation," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2283–2294, 2017.

[10] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[11] S. Li and L. Fang, "Signal denoising with random refined orthogonal matching pursuit," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 1, pp. 26–34, 2012.

[12] S. Li, L. Fang, and H. Yin, "An efficient dictionary learning algorithm and its application to 3-d medical image denoising," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 417–427, 2012.

[13] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsieh, and G. Wang, "Low-dose x-ray ct reconstruction via dictionary learning," *IEEE Trans. Med. Imag.*, vol. 31, no. 9, pp. 1682–1697, 2012.

[14] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, 2012.

[15] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[16] H. M. Nguyen, X. Peng, M. N. Do, and Z.-P. Liang, "Denoising mr spectroscopic imaging data with low-rank approximations," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 78–89, 2013.

[17] Y. Chang, L. Yan, T. Wu, and S. Zhong, "Remote sensing image stripe noise removal: from image decomposition perspective," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7018–7031, 2016.

[18] J. Ye, H. Wang, and W. Yang, "Image recovery for electrical capacitance tomography based on low-rank decomposition," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1751–1759, 2017.

[19] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[20] J. Sun, W. Cao, Z. Xu, J. Ponce *et al.*, "Learning a convolutional neural network for non-uniform motion blur removal." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 769–777.

[21] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[22] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[23] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, 2017.

[24] 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," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, 2017.

[25] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction," *Med. Phys.*, vol. 44, no. 10, 2017.

[26] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *Proc. Int. Conf. Data Mining Workshops*, 2016, pp. 241–246.

[27] J. Wang, Y. Zhao, J. H. Noble, and B. M. Dawant, "Conditional generative adversarial networks for metal artifact reduction in ct images of the ear," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2018, pp. 3–11.

[28] Y. Zhang and H. Yu, "Convolutional neural network based metal artifact reduction in x-ray computed tomography," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1370–1381, 2018.

[29] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau, "Brain mri super-resolution using deep 3d convolutional networks," in *Proc. IEEE Int. Symp. Biomed. Imag.* IEEE, 2017, pp. 197–200.

[30] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, 2018.

[31] D. Wu, K. Kim, G. El Fakhri, and Q. Li, "Iterative low-dose ct reconstruction with priors trained by artificial neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2479–2486, 2017.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[33] P. Gravel, G. Beaudoin, and J. A. De Guise, "A method for modeling noise in medical images," *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1221–1232, 2004.

[34] J. Bresenham, "A linear algorithm for incremental digital display of circular arcs," *Communications of the ACM*, vol. 20, no. 2, pp. 100–106, 1977.

[35] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 479–486.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[38] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[39] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[40] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 624–632.

[41] Y. Chang, L. Yan, H. Fang, S. Zhong, and W. Liao, "Hsi-denet: Hyperspectral image restoration via convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 667–682, 2019.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[43] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[44] F. Luisier, T. Blu, and M. Unser, "A new sure approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 593–606, 2007.

[45] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2862–2869.

[46] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2221–2229, 2009.

[47] L. Zhu, C.-W. Fu, M. S. Brown, and P.-A. Heng, "A non-local low-rank framework for ultrasound speckle reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5650–5658.

[48] P. Wang, H. Zhang, and V. M. Patel, "Sar image despeckling using a convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1763–1767, 2017.

[49] B. Münch, P. Trtik, F. Marone, and M. Stampanoni, "Stripe and ring artifact removal with combined waveletfourier filtering," *Opt. Exp.*, vol. 17, no. 10, pp. 8567–8591, 2009.

[50] J. Fehrenbach, P. Weiss, and C. Lorenzo, "Variational algorithms to remove stationary noise: applications to microscopy imaging," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4420–4430, 2012.

[51] Z. He, Y. Cao, Y. Dong, J. Yang, Y. Cao, and C.-L. Tisse, "Single image based nonuniformity correction of uncooled long-wave infrared detectors: A deep learning approach," *Appl. Opt.*, vol. 57, no. 18, pp. D155–164, 2018.

[52] D. Prell, Y. Kyriakou, and W. A. Kalender, "Comparison of ring artifact correction methods for flat-detector ct," *Phys. Med. Biol.*, vol. 54, no. 12, p. 3881, 2009.

[53] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[54] Y. Chang, L. Yan, H. Fang, and C. Luo, "Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1852–1866, 2015.

[55] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *Proc. IEEE Conf. CVPR*, 2011, pp. 177–184.

[56] W. C. Shu-wen and J.-L. Pellequer, "Destripe: frequency-based algorithm for removing stripe noises from afm images," *BMC Struct. Biol.*, vol. 11, no. 1, p. 7, 2011.

[57] K. H. Jensen, F. J. Sigworth, and S. S. Brandt, "Removal of vesicle structures from transmission electron microscope images," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 540–552, 2016.