2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

NOISE REDUCTION IN LOW-DOSE CT USING A 3D MULTISCALE SPARSE DENOISING AUTOENCODER

Katrin Mentl [⋆,†], *Boris Mailhé* [†], *Florin C. Ghesu* [⋆,†],
Frank Schebesch [⋆], *Tino Haderlein* [⋆], *Andreas Maier* [⋆], *Mariappan S. Nadar* [†]

[⋆]Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen, Germany
katrin.mentl@fau.de
[†]Medical Imaging Technologies, Siemens Healthineers, Princeton NJ, USA
boris.mailhe@siemens-healthineers.com

ABSTRACT

This article presents a novel neural network-based approach for enhancement of 3D medical image data. The proposed networks learn a sparse representation basis by mapping the corrupted input data to corresponding optimal targets. To reinforce the adjustment of the network to the given data, the threshold values are also adaptively learned. In order to capture important image features on various scales and be able to process large computed tomography (CT) volumes in a reasonable time, a multiscale approach is applied. Recursively downsampled versions of the input are used and denoising operator of constant size are learnt at each scale. The networks are trained end-to-end from a database of real high-dose acquisitions with synthetic additional noise to simulate the corresponding low-dose scans. Both 2D and 3D networks are evaluated on CT volumes and compared to the block-matching and 3D filtering (BM3D) algorithm. The presented methods achieve an increase of 4% to 11% in the SSIM and of 2.4 to 2.8 dB in the PSNR with respect to the ground truth, outperform BM3D in quantitative comparisions and present no visible texture artifacts. By exploiting volumetric information, 3D networks achieve superior results over 2D networks.

Index Terms— CT, 3D neural networks, denoising autoencoder

1. INTRODUCTION

CT reconstructs medical images from multiple X-ray projections through the body at multiple orientations. Due to the use of ionizing radiation, CT acquisition must achieve a trade-off between the radiation dose and the signal-to-noise ratio. Therefore, multiple regularized image reconstruction [1, 2] and denoising methods [3, 4] were proposed to reduce the dose required to obtain an intepretable image. Recently, deep learning approaches have shown promising results for denoising [5, 6] and have been applied to CT [7, 8]. However, pure data-driven approaches suffer from an implicit dependency to the acquisition parameters (e.g. the noise level), and deep learning methods can become too expensive when applied to large 3D volumes. Inspired by the work in [9], we propose to find sparse approximations using neural networks that are trained in the fashion of a convolutional sparse denoising autoencoder. The proposed architecture features a pyramidal decomposition to accelerate 3D processing and can accomodate the noise level as separate input to learn how to adapt to it.

2. MULTISCALE SPARSE CODING NETWORKS

We formulate the idea of sparsity-based denoising algorithms in terms of a feed-forward neural network. However, by learning both the dictionary that was randomly initialized as opposed to using predefined appropriate basis filters and the scale of our thresholding function in order to optimally attach the parameters to the given data.

2.1. CT Denoising Problem

Noise texture is always present in CT images: if the image is noise-free, then the patient was arguably subjected to a higher dose than necessary. For that reason, radiologists are trained to read noisy images and may be more comfortable with these than with denoised images. However, image denoising is rarely an end goal on its own. The denoised image is going to be used for some other task and the denoising performance should thus be evaluated (and ideally trained) by its impact on downstream tasks, even if those are not automated. Obtaining natural-looking images using supervised training as proposed in this work provides a significant step towards clinical acceptance.

The feature (mentioned herein) is based on research, and is not commercially available. Due to regulatory reasons its future availability cannot be guaranteed.

978-1-5090-6341-3/17/$31.00 ©2017 IEEE

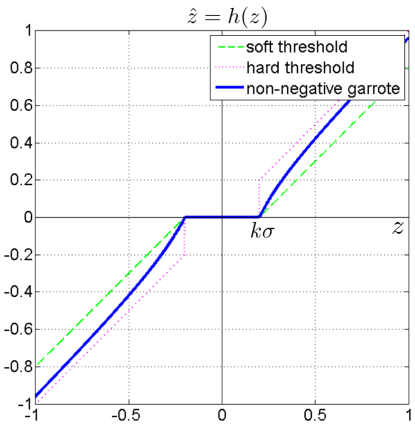

Fig. 1: Plot of the non-negative garrote function (solid) along with the soft thresholding (dotted-dashed) and hard thresholding (dotted) functions.

In general, the image denoising problem consists in estimating a hidden image $\boldsymbol{x}$ as a function of a noisy observation $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$. In CT images reconstructed by filtered backprojection, the noise $\boldsymbol{\epsilon}$ is not white because a low-pass convolution kernel applied during the reconstruction shapes the noise into a texture [10]. Besides, an exact statistical description of the noise in image domain is hard to provide because the noise is non-Gaussian in the raw measurement domain.

2.2. Sparse Denoising Autoencoder

A denoising autoencoder [11] is a neural network $\mathcal{N}$ trained on image pairs $(\boldsymbol{y}, \boldsymbol{x})$ that learns to realise the denoising mapping $\hat{x} \triangleq \mathcal{N}(\boldsymbol{y}) \approx \boldsymbol{x}$. Adopting a supervised learning approach to CT denoising can help in two ways. First, it allows the algorithm to learn the statistics of the noise rather than to use a complex approximate model. Second, if the ground truth images come from a real high-dose clinical, then they still contain some noise, and the algorithm can learn to denoise while preserving the noise texture, which can lead to more natural-looking images and higher perceived image quality [12].

However, learning a noise model can also have the drawback of tying the network to one particular scanner setting. This would make the network hard to use in clinical practice where technologists routinely adjust the dose, e.g. to adapt to the patient's body mass. In the context of sparse denoising, it is well known that one can adapt to the noise level by changing the value of the threshold applied to the obtained coefficients in a certain representation domain [13]. For this reason, this work uses a transform-domain denoiser as the architecture of the autoencoder:

$$\hat{x} = \boldsymbol{W}' h(\boldsymbol{W}\boldsymbol{y}), \tag{1}$$

with $\boldsymbol{W}$ a trainable convolutional decomposition operator, W' a trainable reconstruction operator and h a sparsity-inducing activation function. The number of free parameters is further reduced by imposing $\boldsymbol{W}' = \boldsymbol{W}^T$, which is equivalent to constraining $\boldsymbol{W}$ to be a tight frame.

2.3. Thresholding Function

Sparse denoising algorithms are based on the observation that, given a representation of a noisy image in a suitable basis, the small coefficients will be mainly due to noise while the few large coefficients capture the main image features. Similarly, in our feed-forward neural network an appropriate thresholding function should thus set the small transform coefficients to zero and keep the large coefficients in order to obtain a denoised estimate. The non-negative garrote function [15] was shown to successfully remedy the disadvantages of both soft and hard thresholding functions which are often used for this task [16]. While the soft shrinkage comes with a bigger bias due to the shrinkage of large coefficients, the hard shrinkage function is not continuous and thus more likely to have bigger variance and introduce instability as it is sensitive to small changes in the data. Fig. 1 shows a plot of the non-negative garrote function along with the soft and hard thresholding functions. It is dependent on the noise level σ and a thresholding level k and forces sparsity on each representation coefficient z_j by:

$$\hat{z}_j = h_{\text{garrote}}(z_j) = \frac{(z_j^2 - k\sigma^2)_+}{z_j}, \tag{2}$$

to obtain the thresholded coefficients $\hat{\boldsymbol{z}}$. The positive part function $+$ is defined as $x_+ = \max(x, 0)$. The noise variance σ^2 is an input to the network (which enables training and testing at multiple dose settings), and the thresholding value k is a trainable parameter. Its inital value should be chosen very small to avoid starting in the flat region around 0 where backpropagation would fail to generate gradients.

2.4. Learning a Multiscale Representation

Successful 2D denoising results were reported using filters of size 17×17 [5]. However, such large filters would be prohibitively expensive to apply to large 3D CT volumes of size 512^3 and larger. We rather suggest combining the sparse denoising autoencoder with a pyramidal decomposition as shown in Fig. 2: instead of increasing the filter size, the image is downsampled recursively and processed at all scales with the same filter size. While the operator $\boldsymbol{W}$ could in principle be shared across scales, it was found to degrade the denoising quality in our tests. Therefore, in the presented results 3 distinct operators are learnt.

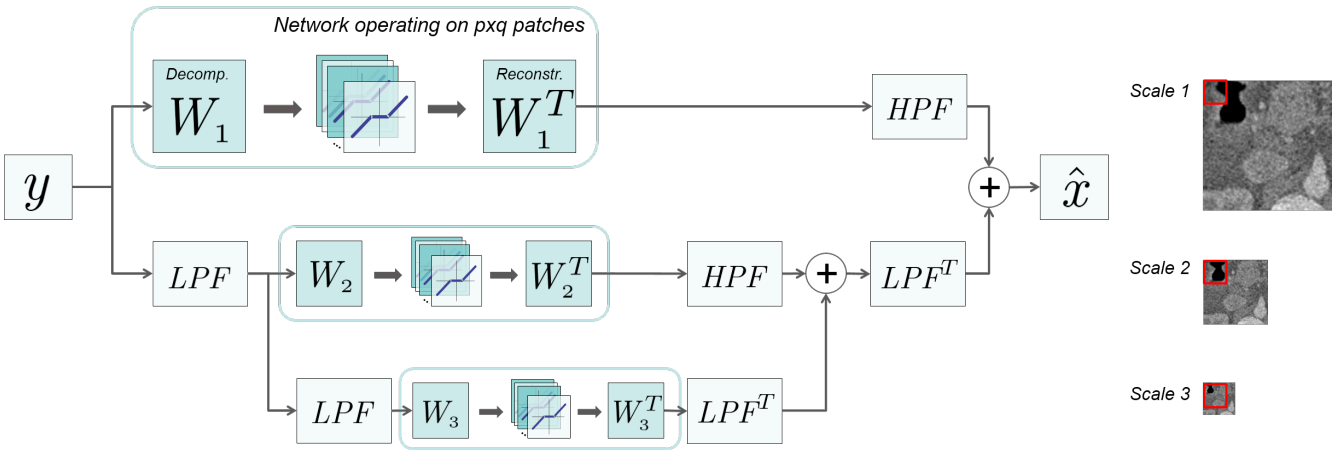

Fig. 2: Block diagram of the multiscale sparse coding network operating on three scale levels. Low-pass wavelet decomposition is carried out by convolution followed by downsampling (denoted by *LPF*) while wavelet reconstruction consists of successive upsampling and transposed convolution (denoted by LPF^T). The high-pass wavelet decomposition and the consecutive reconstruction are both summarized under *HPF*. In absence of the thresholding function, the sum of LPF^T and *HPF* realizes near-perfect reconstruction [14].

2.5. Network Architecture

The architecture of the 2D multiscale sparse coding network is illustrated in Fig. 2. It consists of sparse denoising autoencoder blocks which repeat on further decomposition levels. Each sparse denoising autoencoder first maps its input to a hidden representation by using a convolutional layer which consists of 25 filter kernels of size 5×5 and accordingly of 25 filters of size 5×5×5 in 3D. These 25 filters correspond to the elements of the dictionary $\boldsymbol{W}$ that we seek to learn. The obtained 25 feature maps, i.e. the representation coefficients, are thresholded by applying the non-negative garrote function h_{garrote} (see Equation 2) and eventually reconstructed to the same shape as the autoencoder input by using a transposed convolutional layer with the dictionary filter elements (corresponds to a convolution with a filter of size 25×5×5).

The pyramidal decomposition was implemented with a Daubechies-2 separable orthogonal wavelet basis, using the LL (or LLL in 3D) band as the low-pass filter and summing all other bands into the high-pass part. During decomposition, downsampling is performed together with low-pass filtering by using a strided (stride 2) convolutional layer (summarized under *LPF* in Fig. 2). After denoising, perfect reconstruction of the low-pass branches is achieved by upsampling with zero-filling followed by low-pass filtering LPF^T. On the high-pass branches, the high-pass filter is not applied before the denoising autoencoder so that the images at different scales still have similar contrast. So in order to preserve perfect reconstruction, the high-pass filter block *HPF* applies both the decomposition and reconstruction high-pass wavelet filters.

On each scale level, a distinct operator of the same sized filter kernels $\boldsymbol{W}_1, ..., \boldsymbol{W}_s$ is learnt where s denotes the number of scale levels that are considered. In our network where we apply convolutional layers with filter kernels of size 5×5, the downsampling approach effectively corresponds to processing of the original sized network input with filters of size 10×10 on scale level 2 and with filters of size 20×20 on scale level 3 etc. The 3D network is obtained by replacing all 2D operations with their 3D counterparts.

3. EXPERIMENTS

An effective denoising algorithm should be able to reconstruct sharp edges of organ boundaries and used medical instruments since these often contain the most useful information about the patient's condition. It should not produce artifacts, such as ringing effects along edges, splotchy artifacts or artifical structures, as these impede the physician to make correct decisions. In the following, we evaluate our 2D/3D multiscale sparse coding networks for denoising on synthetically corrupted CT slices/volumes.

3.1. Datasets

The dataset consists of contrast-enhanced CT scans in precontrast, arterial and veinous phases of the liver from 10 different patients with a resolution of 1×1×0.3 mm. From the same raw data, virtual low-dose data were created through the addition of signal-dependent Gaussian noise. The noise addition was calibrated such that the output data pertain to a dose value of 30% of the original dose. These virtual low-dose data were reconstructed to obtain the low-dose volume datasets. This process ensures that the high and low-dose images are perfectly registered, avoids scanning patients twice and guarantees that the noise texture is natural in both images. Since

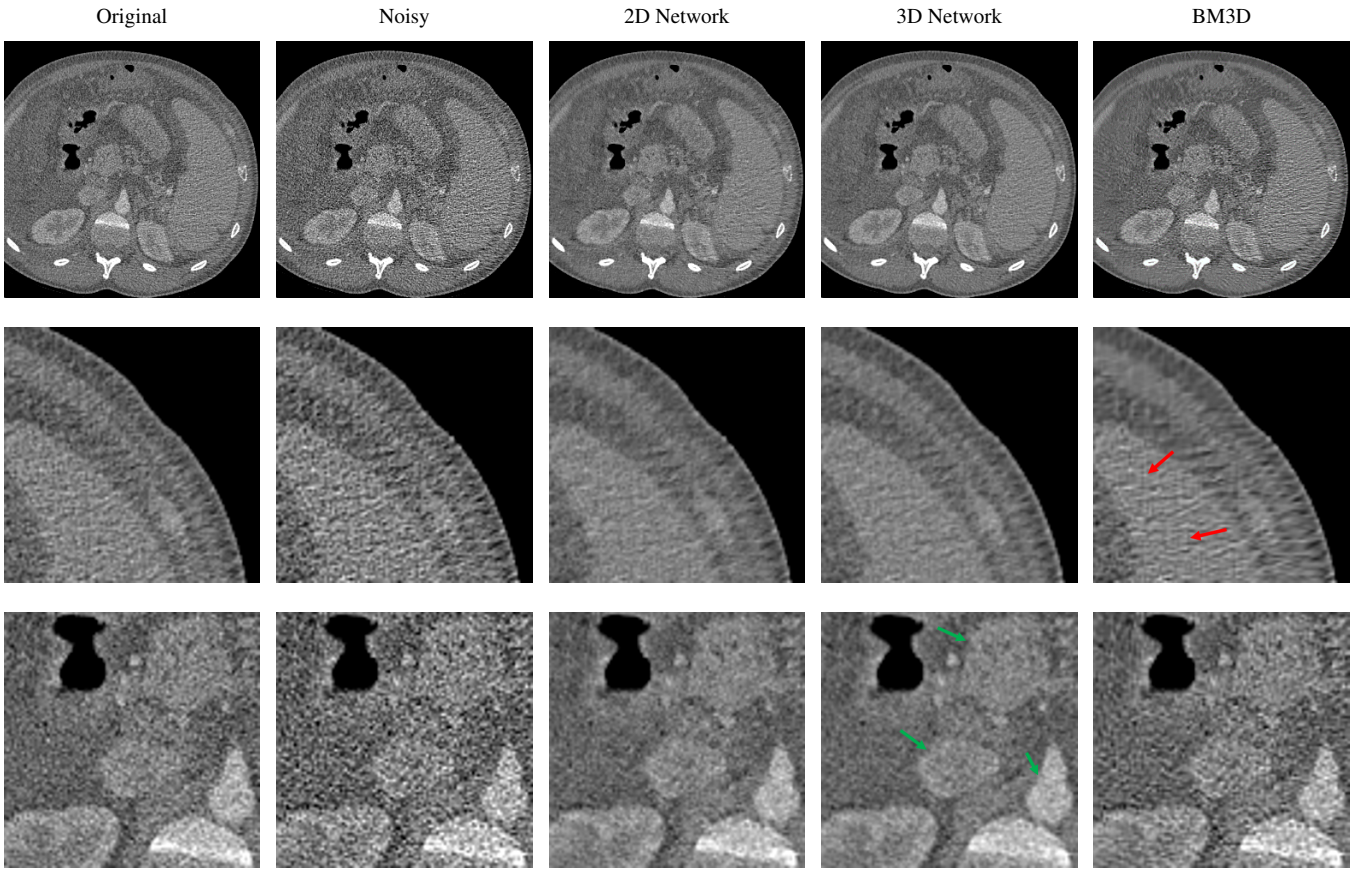

Fig. 3: Comparison of original high-dose image and its corresponding artifically corrupted low-dose image to results obtained with our 2D and 3D networks and the BM3D approach on test dataset 1 whose quantitative results are given in Table 1. The red arrows highlight artifacts produced by the BM3D while the green arrows point to anatomical structures which were recovered by our 3D network and are less visible on the original high-dose acquisition.

training and testing were performed on a single dose reduction setting, the noise level input was not used. Instead, σ was set to 1 and the threshold values were learnt. However, the validity of the approach to train and test at multiple noise levels has already been demonstrated in previous work on 2D images [17]. A split at patient level in 80% - 20% proportion is made to determine the training and validation set. 2D and 3D denoising networks were trained on 150000 overlapping 2D patches extracted from 50 randomly selected CT slices and 10000 overlapping 3D patches extracted from 50 randomly chosen CT volumes, respectively.

3.2. Network and Training Specifications

Both the 3D and 2D networks operate on three scale levels as the use of more than two decomposition levels did not further improve the results. We compare our 3D network, which performs denoising on CT volumes, to a 2D version of the network which is applied to CT slices and eventually to results using the BM3D algorithm [18], which is currently considered state-of-the-art in image denoising. It performs denoising by grouping similar 2D image patches into 3D stacks and subsequently applying collaborative filtering to the similar blocks. Due to the high computational costs of 3D training, the dimensions of our networks were selected post-hoc on the 2D task after conducting several experiments and extrapolated to 3D. In the 2D network, the learned filter basis was selected as 25 filter kernels of size 5×5, and in the 3D case as 25 filter kernels of size 5×5×5. The networks were implemented in Python using the Theano [19] and Lasagne [20] packages. All networks were randomly initialized with Gaussian weights and trained from end-to-end for 1500 epochs with early stopping with the ADAM algorithm [21]. We apply a learning rate of $\mu = 10^{-4}$. The objective function to compute the loss is selected as the l_1 norm of the error [22].

3.3. Results

The 2D and 3D networks were evaluated on CT data using both objective metrics and by visual evaluation. On an Nvidia

Test Dataset	Metrics	Noisy	2D Network	3D Network	BM3D
1	PSNR [dB]	38.54	41.39	41.33	40.57
	SSIM	0.91	0.95	0.95	0.94
2	PSNR [dB]	40.86	43.32	43.36	42.56
	SSIM	0.87	0.97	0.97	0.96

Table 1: Quantitative results (PSNR/SSIM w.r.t. the ground truth) of our 2D/3D networks and the BM3D on two test datasets.

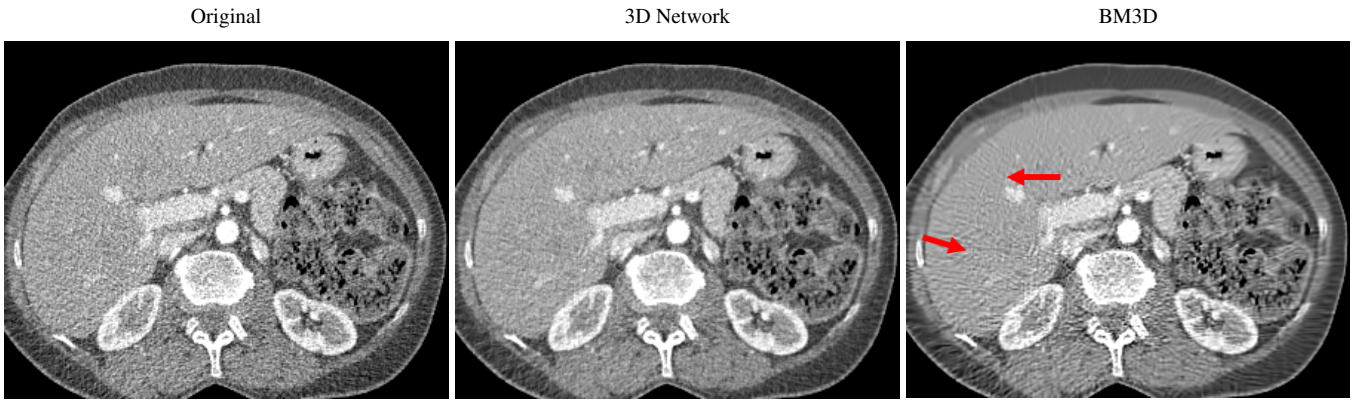

Fig. 4: Comparison of original high-dose image to the 3D network output and the BM3D results on test dataset 2 whose quantitative results are given in Table 1. Texture artifacts produced by the BM3D are clearly visible in homogeneous regions, such as the liver. Some of these regions are highlighted with red arrows.

Geforce GTX GPU, the denoising time for the 3D network is about 1 minute for a 512×512×1000 volume (sliced in multiple 512×512×128 slabs to fit in the GPU memory.)

3.3.1. Quantitative Evaluation.

Quantitative evaluation was performed using the peak-signal-to-noise ratio (PSNR), which is a pixel difference-based measure and the structural similarity index measure (SSIM) [23]. Both the 2D and 3D networks outperform BM3D for both evaluation metrics. For both our test datasets, which we denote as test dataset 1 and test dataset 2, the networks produce outputs with an increased SSIM by around 4% and 11% and an increased PSNR by around 2.8 dB and 2.4 dB, respectively. The SSIM for the 3D network result is equal but unable to exceed the SSIM of the 2D network output, and the PSNR for test dataset is lower for the 3D network than for the 2D network as shown in Table 1.

3.3.2. Qualitative Evaluation and Discussion.

From a visual perspective, our networks avoid producing texture artifacts which are clearly visible in the BM3D results as shown in Fig. 4. The results obtained with the 3D network clearly outperform the 2D-based approach from a visual point of view. By exploiting the rich spatial content, organ boundaries are much sharper reconstructed and small details, such as texture patterns, are better recovered as shown in Fig. 3.

However, the quantitative results that are summarized in Table 1 do not indicate a superior performance of the 3D over the 2D network. One possible explanation is that quantitative evaluation was performed on the full dynamic range images whereas CT images are typically viewed after applying windowing (i.e. dynamic range compression). That choice for quantitative evaluation was made to avoid introducing a bias towards bones or soft tissues. It also explains why the quantitative results are rather high for all methods, including the noisy low-dose image.

4. CONCLUSION

The main contribution of this work is the use of neural networks to learn a sparse representation for 3D data from which noise-free estimates can be reconstructed. Experiments with the CT datasets show that the proposed networks do not introduce noticeable texture artifacts in contrast to the BM3D method. The multiscale networks are able to learn a mapping from artificially corrupted to high-dose data without the need of prior information about underlying noise models of the given data.

Thanks to the introduction of a learnable threshold value that is proportional to the input noise level, the network should be able adapt to the dose. Future works will include testing the efficiency of the method in multiple dose settings, as well as improving its adaptivity to a wider range of imaging parameters, such as the tube voltage (which affects image contrast), region-of-interest, reconstruction resolution, and reconstruction kernel (which affects the noise texture).

5. REFERENCES

[1] S. Gordic, F. Morsbach, B. Schmidt, et al., "Ultralow-dose chest computed tomography for pulmonary nodule detection: first performance evaluation of single energy scanning with spectral shaping," *Investigative Radiology*, vol. 49, no. 7, pp. 465–473, 2014.

[2] Q. Xu, H. Yu, X. Mou, et al., "Low-dose X-ray CT reconstruction via dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 31, no. 9, pp. 1682–1697, 2012.

[3] Z. Li, L. Yu, J. D. Trzasko, et al., "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Medical Physics*, vol. 41, no. 1, 2014, doi:10.1118/1.4851635.

[4] Y. Chen, X. Yin, L. Shi, et al., "Improving abdomen tumor low-dose CT images using a fast dictionary learning based processing," *Physics in Medicine and Biology*, vol. 58, no. 16, pp. 5803, 2013.

[5] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2392–2399.

[6] K. Zhang, W. Zuo, Y. Chen, et al., "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, 2017.

[7] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *arXiv:1610.09736*, 2016.

[8] H. Chen, Y. Zhang, W. Zhang, et al., "Low-dose CT via convolutional neural network," *Biomedical Optics Express*, vol. 8, no. 2, pp. 679–694, 2017.

[9] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 399–406.

[10] J.A. Fessler, "Fundamentals of CT reconstruction in 2D and 3D," in *Comprehensive Biomedical Physics*, Anders Brahme, Ed., pp. 263–295. Elsevier, Oxford, 2014.

[11] P. Vincent, H. Larochelle, Y. Bengio, et al., "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International conference on Machine learning (ICML)*, 2008, pp. 1096–1103.

[12] E. C. Ehman, L. Yu, A. Manduca, et al., "Methods for clinical evaluation of noise reduction techniques in abdominopelvic CT," *RadioGraphics*, vol. 34, no. 4, pp. 849–862, 2014.

[13] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[14] I. Daubechies, "Ten lectures on wavelets," *Society for Industrial and Applied Mathematics (SIAM)*, vol. 1, 1992.

[15] Leo Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[16] Hong-Ye Gao, "Wavelet shrinkage denoising using the non-negative garrote," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 469–488, 1998.

[17] Y. Matviychuk, B. Mailhé, X. Chen, et al., "Learning a multiscale patch-based representation for image denoising in X-ray fluoroscopy," in *International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2330–2334.

[18] K. Dabov, A. Foi, V. Katkovnik, et al., "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[19] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv:1605.02688*, 2016.

[20] S. Dieleman, J. Schlúter, C. Raffel, et al., "Lasagne: First release.," Aug. 2015.

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

[22] H. Zhao, O. Gallo, I. Frosio, et al., "Is L2 a good loss function for neural networks for image processing?," *arXiv:1511.08861*, 2015.

[23] Z. Wang, A. C. Bovik, H. Sheikh, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.