Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss

Qingsong Yang, Pingkun Yan, *Senior Member, IEEE*, Yanbo Zhang, *Member, IEEE*, Hengyong Yu, *Senior Member, IEEE*, Yongyi Shi, Xuanqin Mou, *Senior Member, IEEE*, Mannudeep K. Kalra, Yi Zhang, *Member, IEEE*, Ling Sun, and Ge Wang, *Fellow, IEEE*

***Abstract*— The continuous development and extensive use of computed tomography (CT) in medical practice has raised a public concern over the associated radiation dose to the patient. Reducing the radiation dose may lead to increased noise and artifacts, which can adversely affect the radiologists' judgment and confidence. Hence, advanced image reconstruction from low-dose CT data is needed to improve the diagnostic performance, which is a challenging problem due to its ill-posed nature. Over the past years, various low-dose CT methods have produced impressive results. However, most of the algorithms developed for this application, including the recently popularized deep learning techniques, aim for minimizing the mean-squared error (MSE) between a denoised CT image and the ground truth under generic penalties. Although the peak signal-to-noise ratio is improved, MSE- or weighted-MSE-based methods can compromise the visibility of important structural details after aggressive denoising. This paper introduces a new CT image denoising method based on the generative adversarial network (GAN) with Wasserstein distance and perceptual similarity. The Wasserstein distance is a key concept of the optimal transport theory and promises to improve the performance of GAN. The perceptual loss suppresses noise by comparing the perceptual features of a denoised output against those of the ground truth in an established feature space, while the GAN focuses more on migrating the data noise distribution from strong to weak statistically. Therefore, our proposed method transfers our knowledge of visual perception to the image denoising task and is capable of not only reducing the image noise level but also trying to keep the critical information at the same time. Promising results have been obtained in our experiments with clinical CT images.**

***Index Terms*— Low dose CT, image denoising, deep learning, perceptual loss, WGAN.**

I. Introduction

X-RAY computed tomography (CT) is one of the most important imaging modalities in modern hospitals and clinics. However, there is a potential radiation risk to the patient, since x-rays could cause genetic damage and induce cancer in a probability related to the radiation dose [1], [2]. Lowering the radiation dose increases the noise and artifacts in reconstructed images, which can compromise diagnostic information. Hence, extensive efforts have been made to design better image reconstruction or image processing methods for low-dose CT (LDCT). These methods generally fall into three categories: (a) sinogram filtration before reconstruction [3]–[5], (b) iterative reconstruction [6], [7], and (c) image post-processing after reconstruction [8]–[10].

Over the past decade, researchers were dedicated to developing new iterative algorithms (IR) for LDCT image reconstruction. Generally, those algorithms optimize an objective function that incorporates an accurate system model [11], [12], a statistical noise model [13]–[15] and prior information in the image domain. Popular image priors include total variation (TV) and its variants [16]–[18], as well as dictionary learning [19], [20]. These iterative reconstruction algorithms greatly improved image quality but they may still lose some details and suffer from remaining artifacts. Also, they require a high computational cost, which is a bottleneck in practical applications.

Manuscript received December 20, 2017; revised February 21, 2018 and March 26, 2018; accepted April 10, 2018. Date of publication April 17, 2018; date of current version May 31, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671312 and in part by the National Institute of Biomedical Imaging and Bioengineering/National Institutes of Health under Grant R01 EB016977 and Grant U01 EB017140. *(Corresponding author: Pingkun Yan.)*

Q. Yang, P. Yan, and G. Wang are with the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: yangq4@rpi.edu; yanp2@rpi.edu; wangg6@rpi.edu).

Y. Zhang and H. Yu are with the Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854 USA (e-mail: yanbo_zhang@uml.edu; hengyong-yu@ieee.org).

Y. Shi and X. Mou are with the Institute of Image Processing and Pattern Recognition, Xian Jiaotong University, Xian 710049, China (e-mail: xqmou@mail.xjtu.edu.cn).

M. K. Kalra is with the Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114 USA (e-mail: mkalra@mgh.harvard.edu).

Y. Zhang is with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yzhang@scu.edu.cn).

L. Sun is with the Huaxi MR Research Center, Department of Radiology, West China Hospital, Sichuan University, Chengdu 610041, China (e-mail: 251834489@qq.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2827462

On the other hand, sinogram pre-filtration and image post-processing are computationally efficient compared to iterative reconstruction. Noise characteristic was well modeled in the sinogram domain for sinogram-domain filtration. However, sinogram data of commercial scanners are not readily available to users, and these methods may suffer from resolution loss and edge blurring. Sinogram data need to be carefully processed, otherwise artifacts may be induced in the reconstructed images.

Differently from sinogram denoising, image post-processing directly operates on an image. Many efforts were made in the image domain to reduce LDCT noise and suppress artifacts. For example, the non-local means (NLM) method was adapted for CT image denoising [8]. Inspired by compressed sensing methods, an adapted K-SVD method was proposed [9] to reduce artifacts in CT images. The block-matching 3D (BM3D) algorithm was used for image restoration in several CT imaging tasks [10], [21]. With such image post-processing, image quality improvement was clear but over-smoothing and/or residual errors were often observed in the processed images. These issues are difficult to address, given the non-uniform distribution of CT image noise.

The recent explosive development of deep neural networks suggests new thinking and huge potential for the medical imaging field [22], [23]. As an example, the LDCT denoising problem can be solved using deep learning techniques. Specifically, the convolutional neural network (CNN) for image super-resolution [24] was recently adapted for low-dose CT image denoising [25], with a significant performance gain. Then, more complex networks were proposed to handle the LDCT denoising problem such as the RED-CNN in [26] and the wavelet network in [27]. The wavelet network adopted the shortcut connections introducted by the U-net [28] directly and the RED-CNN [27] replaced the pooling/unpooling layers of U-net with convolution/deconvolution pairs.

Despite the impressive denoising results with these innovative network structures, they fall into a category of an end-to-end network that typically uses the mean squared error (MSE) between the network output and the ground truth as the loss function. As revealed by [29] and [30], this per-pixel MSE is often associated with over-smoothed edges and loss of details. As an algorithm tries to minimize per-pixel MSE, it overlooks subtle image textures/signatures critical for human perception. It is reasonable to assume that CT images distribute over some manifolds. From that point of view, the MSE based approach tends to take the mean of high-resolution patches using the Euclidean distance rather than the geodesic distance. Therefore, in addition to the blurring effect, artifacts are also possible such as non-uniform biases.

To tackle the above problems, here we propose to use a generative adversarial network (WGAN) [31] with the Wasserstein distance as the discrepancy measure between distributions and a perceptual loss that computes the difference between images in an established feature space [29], [30].

The use of WGAN is to encourage that denoised CT images share the same distribution as that of normal dose CT (NDCT) images. In the GAN framework, a generative network G and a discriminator network D are coupled tightly and trained simultaneously. While the G network is trained to produce realistic images $G(z)$ from a random vector z, the D network is trained to discriminate between real and generated images [32], [33]. GANs have been used in many applications such as single image super-resolution [29], art creation [34], [35], and image transformation [36]. In the field of medical imaging, Nie *et al.* [37] proposed to use GAN to estimate CT image from its corresponding MR image. Wolterink *et al.* [38] are the first to apply GAN network for cardiac CT image denoising. And Yu *et al.* [39] used GAN network to handle the de-alising problem for fast CS-MRI. Promising results were achieved in these works. We will discuss and compare the results of those two networks in Section III since the proposed network is closely related with their works.

Despite its success in these areas, GANs still suffer from a remarkable difficulty in training [33], [40]. In the original GAN [32], D and G are trained by solving the following minimax problem

$$\min_G \max_D L_{\mathrm{GAN}}(D, G) = \mathbb{E}_{\boldsymbol{x}\sim P_r}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim P_z}[\log(1 - D(G(\boldsymbol{z})))] \quad (1)$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator; P_r and P_z are the real data distribution and the noisy data distribution. The generator G transforms a noisy sample to mimic a real sample, which defines a data distribution, denoted by P_g. When D is trained to become an optimal discriminator for a fixed G, the minimization search for G is equivalent to minimizing the Jensen-Shannon (JS) divergence of P_r and P_g, which will lead to vanished gradient on the generator G [40] and G will stop updating as the training continues.

Consequently, Arjovsky *et al.* [31] proposed to use the *Earth-Mover* (EM) distance or Wasserstein metric between the generated image samples and real data for GAN, which is referred to as WGAN, because the EM distance is continuous and differentiable almost everywhere under some mild assumptions while neither KL nor JS divergence is. After that, an improved WGAN with *gradient penalty* was proposed [41] to accelerate the convergence.

The rationale behind the perceptual loss is two-fold. First, when a person compares two images, the perception is not performed pixel-by-pixel. Human vision actually extracts and compares features from images [42]. Therefore, instead of using pixel-wise MSE, we employ another pre-trained deep CNN (the famous VGG [43]) for feature extraction and compare the denoised output against the ground truth in terms of the extracted features. Second, from a mathematical point of view, CT images are not uniformly distributed in a high-dimensional Euclidean space. They reside more likely in a low-dimensional manifold. With MSE, we are not measuring the intrinsic similarity between the images, but just their superficial differences in the brute-force Euclidean distance. By comparing images according their intrinsic structures, we should project them onto a manifold and calculate the geodesic distance instead. Therefore, the use of the perceptual loss for WGAN should facilitate producing results with not only lower noise but also sharper details.

In particular, we treat the LDCT denoising problem as a transformation from LDCT to NDCT images. WGAN provides a good distance estimation between the denoised LDCT and NDCT image distributions. Meanwhile, the VGG-based perceptual loss tends to keep the image content after denoising. The rest of this paper is organized as follows. The proposed method is described in Section II. The experiments and results are presented in Section III. Finally, relevant issues are discussed and a conclusion is drawn in Section IV.

II. Methods

A. Noise Reduction Model

Let $\boldsymbol{z} \in \mathbb{R}^{N\times N}$ denote a LDCT image and $\boldsymbol{x} \in \mathbb{R}^{N\times N}$ denote the corresponding NDCT image. The goal of the denoising process is to seek a function G that maps LDCT $\boldsymbol{z}$ to NDCT $\boldsymbol{x}$:

$$G : z \rightarrow \boldsymbol{x} \tag{2}$$

On the other hand, we can also take $\boldsymbol{z}$ as a sample from the LDCT image distribution P_L and $\boldsymbol{x}$ from the NDCT distribution or the real distribution P_r. The denoising function G maps samples from P_L into a certain distribution P_g. By varying the function G, we aim to change P_g to make it close to P_r. In this way, we treat the denoising operator as moving one data distribution to another.

Typically, noise in x-ray photon measurements can be simply modeled as the combination of Poisson quantum noise and Gaussian electronic noise. On the contrary, in the reconstructed images, the noise model is usually complicated and non-uniformly distributed across the whole image. Thus there is no clear clue that indicates how data distributions of NDCT and LDCT images are related to each other, which makes it difficult to denoise LDCT images using traditional methods. However, this uncertainty of noise model can be ignored in deep learning denoising because a deep neural network itself can efficiently learn high-level features and a representation of data distribution from modest sized image patches through a neural network.

B. WGAN

Compared to the original GAN network, WGAN uses the Wasserstein distance instead of the JS divergence to compare data distributions. It solves the following minimax problem to obtain both D and G [41]:

$$\min_G \max_D L_{\text{WGAN}}(D, G) = -\mathbb{E}_{\boldsymbol{x}}[D(\boldsymbol{x})] + \mathbb{E}_z[D(G(z))] + \lambda \mathbb{E}_{\hat{\boldsymbol{x}}}[(||\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})||_2 - 1)^2], \tag{3}$$

where the first two terms perform a Wasserstein distance estimation; the last term is the gradient penalty term for network regularization; $\hat{\boldsymbol{x}}$ is uniformly sampled along straight lines connecting pairs of generated and real samples; and λ is a constant weighting parameter. Compared to the original GAN, WGAN removes the log function in the losses and also drops the last sigmoid layer in the implementation of the discriminator D. Specifically, the networks D and G are trained alternatively by fixing one and updating the other.

C. Perceptual Loss

While the WGAN network encourages that the generator transforms the data distribution from high noise to a low noise version, another part of the loss function is added for the network to keep image details or information content. Typically, a mean squared error (MSE) loss function is used, which tries to minimize the pixel-wise error between a denoised patch $G(\boldsymbol{z})$ and a NDCT image patch $\boldsymbol{x}$ as [25], [26]

$$L_{\text{MSE}}(G) = \mathbb{E}_{(\boldsymbol{x},z)}\left[\frac{1}{N^2}||G(\boldsymbol{z}) - \boldsymbol{x}||_F^2\right], \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm. However, the MSE loss can potentially generate blurry images and cause the distortion or loss of details. Thus, instead of using a MSE measure, we apply a perceptual loss function defined in a feature space

$$L_{\text{Perceptual}}(G) = \mathbb{E}_{(\boldsymbol{x},z)}\left[\frac{1}{whd}||\phi(G(\boldsymbol{z})) - \phi(\boldsymbol{x})||_F^2\right], \tag{5}$$

where ϕ is a feature extractor, and w, h, and d stand for the width, height and depth of the feature space, respectively. In our implementation, we adopt the well-known pre-trained VGG-19 network [43] as the feature extractor. Since the pre-trained VGG network takes color images as input while CT images are in grayscale, we duplicated the CT images to make RGB channels before they are fed into the VGG network. The VGG-19 network contains 16 convolutional layers followed by 3 fully-connected layers. The output of the 16th convolutional layer is the feature extracted by the VGG network and used in the perceptual loss function,

$$L_{\text{VGG}}(G) = \mathbb{E}_{(\boldsymbol{x},z)}\left[\frac{1}{whd}||VGG(G(\boldsymbol{z})) - VGG(\boldsymbol{x})||_F^2\right] \tag{6}$$

For convenience, we call the perceptual loss computed by VGG network *VGG loss*.

Combining Eqs. (3) and (6) together, we get the overall joint loss function expressed as

$$\min_G \max_D L_{\text{WGAN}}(D, G) + \lambda_1 \, L_{\text{VGG}}(G) \tag{7}$$

where λ_1 is a weighting parameter to control the trade-off between the WGAN adversarial loss and the VGG perceptual loss.

D. Network Structures

The overall view of the proposed network structure is shown in Fig. 1. For convenience, we name this network WGAN-VGG. It consists three parts. The first part is the generator G, which is a convolutional neural network (CNN) of 8 convolutional layers. Following the common practice in the deep learning community [44], small 3×3 kernels were used in each convolutional layer. Due to the stacking structure, such a network can cover a large enough receptive field efficiently. Each of the first 7 hidden layers of G have 32 filters. The last layer generates only one feature map with a single 3×3 filter, which is also the output of G. We use Rectified Linear Unit (ReLU) as the activation function.

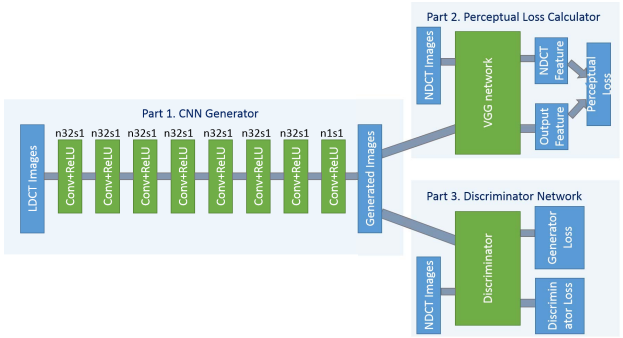


Fig. 1. The overall structure of the proposed WGAN-VGG network. In Part 1, *n* stands for the number of convolutional kernels and *s* for convolutional stride. So, *n*32*s*1 means the convolutional layer has 32 kernels with stride 1.

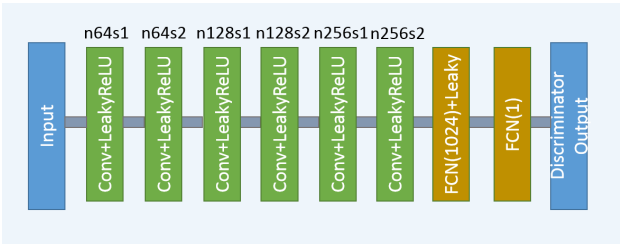


Fig. 2. The structure of the discriminator network. *n* and *s* have the same meaning as in Fig. 1.

The second part of the network is the perceptual loss calculator, which is realized by the pre-trained VGG network [43]. A denoised output image $G(z)$ from the generator G and the ground truth image $\boldsymbol{x}$ are fed into the pre-trained VGG network for feature extraction. Then, the objective loss is computed using the extracted features from a specified layer according to Eq. (6). The reconstruction error is then back-propagated to update the weights of G only, while keeping the VGG parameters intact.

The third part of the network is the discriminator D. As shown in Fig. 2, D has 6 convolutional layers with the structure inspired by Johnson *et al.* [29], Ledig *et al.* [30], and Simonyan and Zisserman [43]. The first two convolutional layers have 64 filters, then followed by two convolutional layers of 128 filters, and the last two convolutional layers have 256 filters. Following the same logic as in G, all the convolutional layers in D have a small 3×3 kernel size. After the six convolutional layers, there are two fully-connected layers, of which the first has 1024 outputs and the other has a single output. Following the practice in [31], there is no sigmoid cross entropy layer at the end of D.

The network is trained using image patches and applied on entire images. The details are provided in Section III on experiments.

E. Other Networks

For comparison, we also trained four other networks.

- CNN-MSE with only MSE loss
- CNN-VGG with only VGG loss
- WGAN-MSE with MSE loss in the WGAN framework
- WGAN with no other additive losses
- Original GAN

All the trained networks are summarized in Table. I.

TABLE I
SUMMARY OF ALL TRAINED NETWORKS: THEIR LOSS FUNCTIONS AND TRAINABLE NETWORKS

Network	Loss
CNN-MSE	$\min_G L_{\mathrm{MSE}}(G)$
WGAN-MSE	$\min_G \max_D L_{\mathrm{WGAN}}(G,D) + \lambda_2 L_{\mathrm{MSE}}(G)$
CNN-VGG	$\min_G L_{\mathrm{VGG}}(G)$
WGAN-VGG	$\min_G \max_D L_{\mathrm{WGAN}}(G,D) + \lambda_1 L_{\mathrm{VGG}}(G)$
WGAN	$\min_G \max_D L_{\mathrm{WGAN}}(G,D)$
GAN	$\min_G \max_D L_{\mathrm{GAN}}(G,D)$

III. EXPERIMENTS

A. Experimental Datasets

We used a real clinical dataset authorized for *"the 2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge"* by Mayo Clinic for the training and evaluation of the proposed networks [45]. The dataset contains 10 anonymous patients' normal-dose abdominal CT images and simulated quarter-dose CT images. In our experiments, we randomly extracted 100,096 pairs of image patches from 4,000 CT images as our training inputs and labels. The patch size is 64×64. Also, we extracted 5,056 pairs of patches from another 2,000 images for validation. When choosing the image patches, we excluded image patches that were mostly air. For comparison, we implemented a state-of-the-art 3D dictionary learning reconstruction technique as a representative IR algorithm [19], [20]. The dictionary learning reconstruction was performed from the LDCT projection data provided by Mayo Clinic.

B. Network Training

In our experiments, all the networks were optimized using Adam algorithm [46]. The optimization procedure for WGAN-VGG network is shown in Fig. 3. The mini-batch size was 128. The hyper-parameters for Adam were set as $\alpha = 1 \times 10^{-5}, \beta_1 = 0.5, \beta_2 = 0.9$, and we chose $\lambda = 10$ as suggested in [41], $\lambda_1 = 0.1, \lambda_2 = 0.1$ according to our experimental experience. The optimization processes for WGAN-MSE and WGAN are similar except that line 12 was changed to the corresponding loss function, and for CNN-MSE and CNN-VGG, lines 2-10 were removed and line 12 was changed according to their loss functions.

The networks were implemented in Python with the Tensorflow library [47]. A NVIDIA Titan XP GPU was used in this study.

C. Network Convergence

To visualize the convergence of the networks, we calculated the MSE loss and VGG loss over the 5,056 image patches for validation according to Eqs. (4) and (6) after each epoch. Fig. 4 shows the averaged MSE and VGG losses respectively versus the number of epochs for the five networks. Even though these two loss functions were not used at the same time for a given network, we still want to see how their values change during the training. In the two figures, both the MSE and VGG losses decreased initially, which indicates that the two

Require: Set hyper-parameters, $\lambda = 10, \alpha = 1 \times 10^{-5}, \beta_1 = 0.5, \beta_2 = 0.9, \lambda_1 = 0.1, \lambda_2 = 0.1$,
Require: Set the number of total epochs, $N_{epoch} = 100$, the number of iteration for discriminator training, $N_D = 4$, the batch size $m = 128$, and image patch size of 80×80.
Require: Initial discriminator parameters w_0, initial generator parameters θ_0
Require: Load VGG-19 network parameters
for $num\_epoch = 0, ..., N_{epoch}$ **do**
  for $t = 1, ..., N_D$ **do**
    Sample a batch of NDCT image patches $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m}$, latent LDCT patches $\{\boldsymbol{z}^{(i)}\}_{i=1}^{m}$, and random numbers $\{\epsilon^{(i)}\}_{i=1}^{m} \sim \text{Uniform}[0, 1]$
    for $i = 1, ..., m$ **do**
      $\hat{\boldsymbol{x}}^{(i)} \leftarrow \epsilon^{(i)}\boldsymbol{x}^{(i)} + (1 - \epsilon^{(i)})G(\boldsymbol{z}^{(i)})$
      $L^{(i)}(D) \leftarrow D(G(\boldsymbol{z}^{(i)})) - D(\boldsymbol{x}^{(i)}) + \lambda(||\nabla D(\hat{\boldsymbol{x}}^{(i)})||_2 - 1)^2$
    end for
  end for
  Update D: $w \leftarrow \text{Adam}(\nabla_w \frac{1}{m}\sum_{i=1}^{m} L^{(i)}(D), w, \alpha, \beta_1, \beta_2)$
  Sample a batch of LDCT patches $\{\boldsymbol{z}^{(i)}\}_{i=1}^{m}$ and corresponding NDCT patches $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m}$,
  for $i = 1, ..., m$ **do**
    $L^{(i)}(G) \leftarrow \lambda_1 L_{\text{VGG}}(\boldsymbol{z}^{(i)}, \boldsymbol{x}^{(i)}) - D(G(\boldsymbol{z}^{(i)}))$
  end for
  Update $G, \theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} L^{(i)}(G), \theta, \alpha, \beta_1, \beta_2)$
end for

Fig. 3. Optimization procedure of WGAN-VGG network.

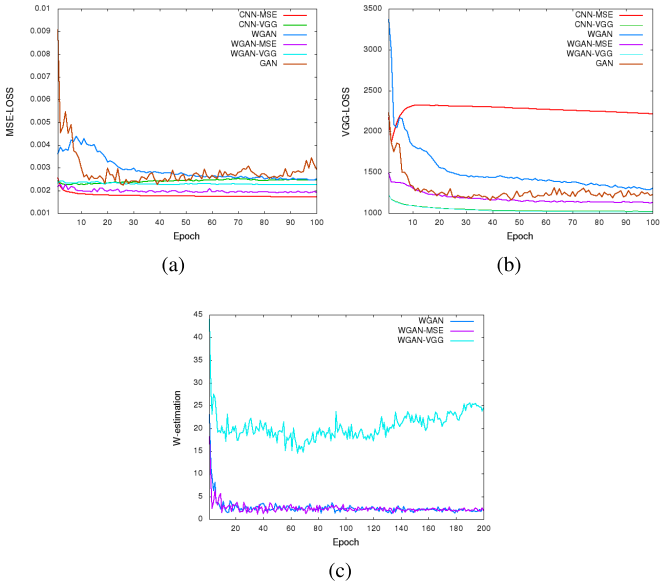


Fig. 4. Plots of validation loss versus the number of epochs during the training of the 5 networks. (a) MSE loss convergence, (b) VGG loss convergence and (c) Wasserstein estimation convergence.

metrics are positively correlated. However, the loss values of the networks in terms of MSE are increasing in the following order, CNN-MSE<WGAN-MSE<WGAN-VGG<CNN-VGG (Fig. 4a), yet the VGG loss are in the opposite order (Fig. 4b). The MSE and VGG losses of GAN network are oscillating in the converging process. WGAN-VGG and CNN-VGG have very close VGG loss values, while their MSE losses are

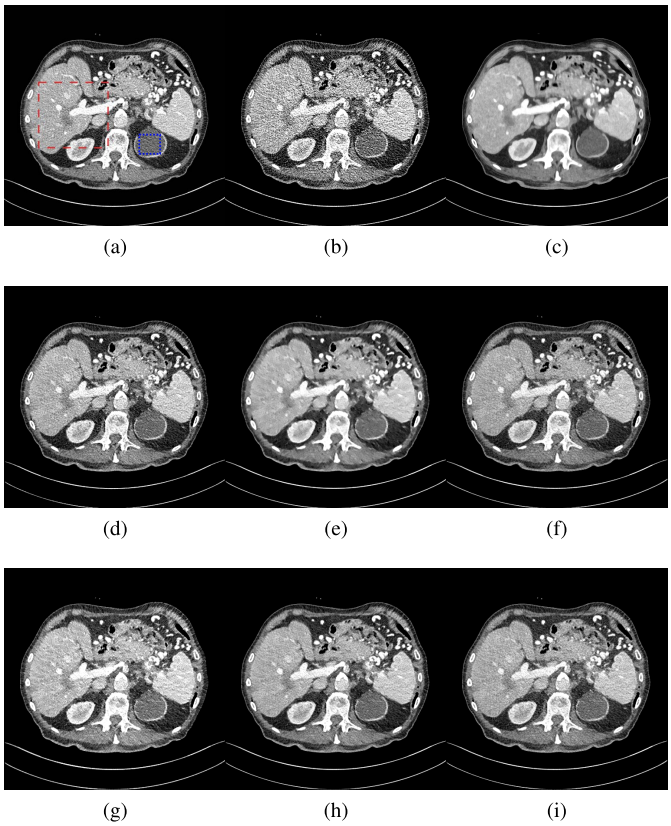


Fig. 5. Transverse CT images of the abdomen demonstrate a low attenuation liver lesion (in the red box) and a cystic lesion in the upper pole of the left kidney (in the blue box). This display window is [−160, 240]HU. (a) Full Dose FBP. (b) Quarter Dose FBP. (c) DictRecon. (d) GAN. (e) CNN-MSE. (f) CNN-VGG. (g) WGAN. (h) WGAN-MSE. (i) WGAN-VGG.

quite different. On the other hand, WGAN perturbed the convergence as measured by MSE but smoothly converged in terms of VGG loss. These observations suggest that the two metrics have different focuses when being used by the networks. The difference between MSE and VGG losses will be further revealed in the output images of the generators.

In order to show the convergence of WGAN part, we plotted the estimated Wasserstein values defined as $|-\mathbb{E}[D(\boldsymbol{x})] + \mathbb{E}[D(G(\boldsymbol{z}))]|$ in Eq. (3). It can be observed in Fig. 4(c) that increasing the number of epochs did reduce the W-distance, although the decay rate becomes smaller. For the WGAN-VGG curve, the introduction of VGG loss has helped to improve the perception/visibility at a cost of a compromised loss measure. For the WGAN and WGAN-MSE curves, we would like to note that what we computed is a surrogate for the W-distance which has not been normalized by the total number of pixels, and if we had done such a normalization the curves would have gone down closely to zero after 100 epochs.

D. Denoising Results

To show the denoising effect of the selected networks, we took two representative slices as shown in Figs. 5 and 7. And Figs. 6 and 8 are the zoomed regions-of-interest (ROIs) marked by the red rectangles in Figs. 5 and 7. All the networks demonstrated certain denoising capabilities.

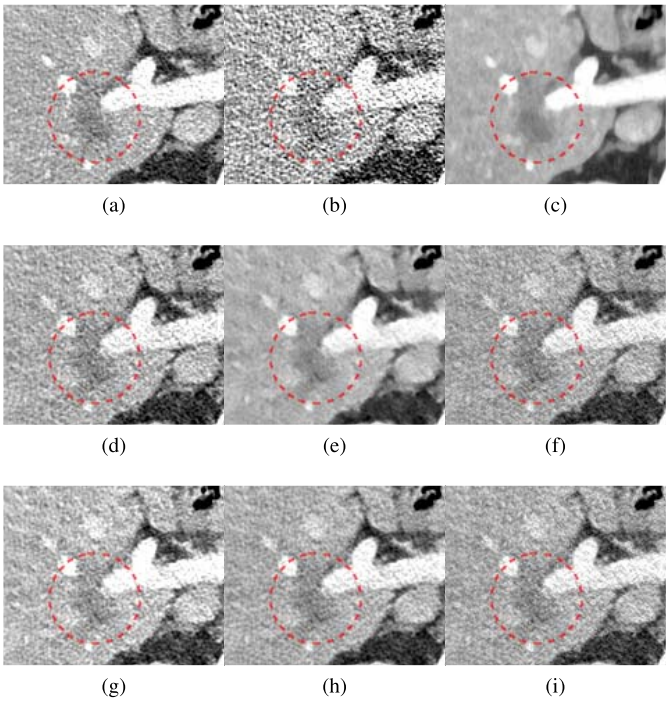


Fig. 6. Zoomed ROI of the red rectangle in Fig. 5. The low attenuation liver lesion with in the dashed circle represents metastasis. The lesion is difficult to assess on quarter dose FBP recon (b) due to high noise content. This display window is [−160, 240]HU. (a) Full Dose FBP. (b) Quarter Dose FBP. (c) DictRecon. (d) GAN. (e) CNN-MSE. (f) CNN-VGG. (g) WGAN. (h) WGAN-MSE. (i) WGAN-VGG

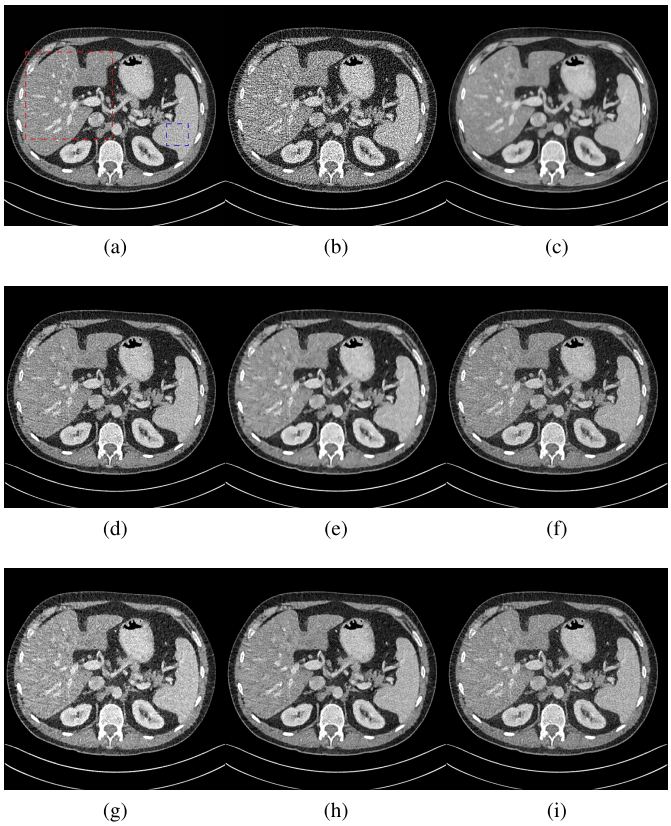


Fig. 7. Transverse CT images of the abdomen demonstrate small low attenuation liver lesions. The display window is [−160, 240]HU. (a) Full Dose FBP. (b) Quarter Dose FBP. (c) DictRecon. (d) GAN. (e) CNN-MSE. (f) CNN-VGG. (g) WGAN. (h) WGAN-MSE. (i) WGAN-VGG.

However, CNN-MSE blurred the images and introduced waxy artifacts as expected, which are easily observed in the zoomed ROIs in Figs. 6e and 8e. WGAN-MSE was able to improve the result of CNN-MSE by avoiding over-smooth but minor streak artifacts can still be observed especially compared to CNN-VGG and WGAN-VGG. Meanwhile, using WGAN or GAN alone generated stronger noise (Figs. 6g and 8g) than the other networks enhanced a few white structures in the WGAN/GAN generated images, which are originated from the low dose streak artifact in LDCT images, while on the contrary the CNN-VGG and WGAN-VGG images are visually more similar to the NDCT images. This is because the VGG loss used in CNN-VGG and WGAN-VGG is computed in a feature space that is trained previously on a very large natural image dataset [48]. By using VGG loss, we transferred the knowledge of human perception that is embedded in VGG network to CT image quality evaluation. The performance of using WGAN or GAN alone is not acceptable because it only maps the data distribution from LDCT to NDCT but does not guarantee the image content correspondence. As for the lesion detection in these two slices, all the networks enhance the lesion visibility compared to the original noisy low dose FBP images as noise is reduced by the different approaches.

As for iterative reconstruction technique, the reconstruction results depend greatly on the choices of the regularization parameters. The implemented dictionary learning reconstruction (DictRecon) result gave the most aggressive noise reduction effect compared to the network outputs as a result of strong regularization. However, it over-smoothed some fine structures. For example, in Fig. 8, the vessel pointed by the green arrow was smeared out while it is easily identifiable in NDCT as well as WGAN-VGG images. Yet, as an iterative reconstruction method, DictRecon has its advantage over post-processing method. As pointed by the red arrow in Fig 8, there is a bright spot which can be seen in DictRecon and NDCT images, but is not observable in LDCT and network processed images. Since the WGAN-VGG image is generated from LDCT image, in which this bright spot is not easily observed, it is reasonable that we do not see the bright spot in the images processed by neural networks. In other words, we do not want the network to generate structure that does not exist in the original images. In short, the proposed WGAN-VGG network is a post-processing method and information that is lost during the FBP reconstruction cannot easily be recovered, which is one limitation for all the post-processing methods. On the other hand, as an iterative reconstruction method, DictRecon algorithm generates images from raw data, which has more information than the post-processing methods.

E. Quantitative Analysis

For quantitative analysis, we calculated the peak-to-noise ratio (PSNR) and structural similarity (SSIM). The summary data are in Table II. CNN-MSE ranks the first in terms of PSNR, while WGAN is the worst. Since PSNR is equivalent

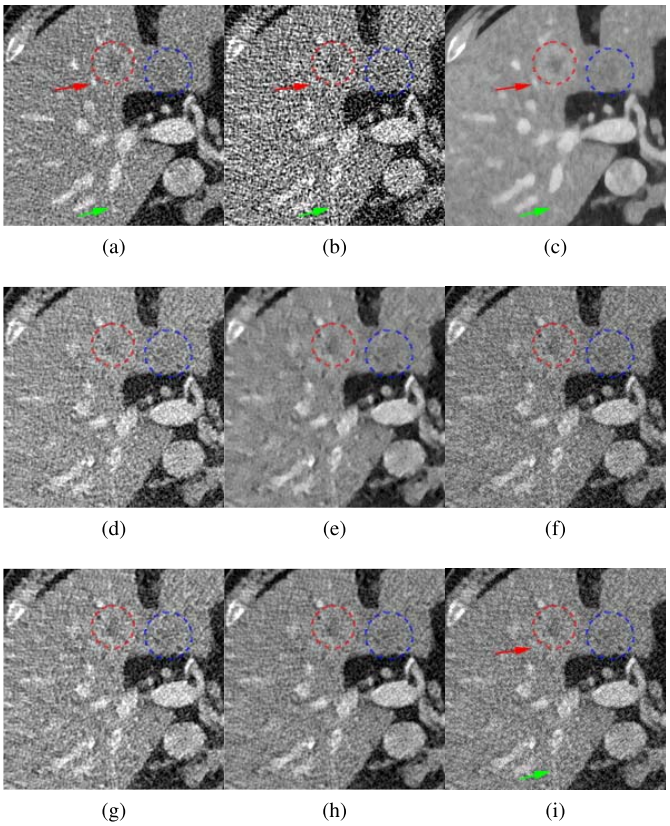


Fig. 8. Zoomed ROI of the red rectangle in Fig. 7 demonstrates the two attenuation liver lesions in the red and blue circles. The display window is [−160, 240]HU. (a) Full Dose FBP. (b) Quarter Dose FBP. (c) DictRecon. (d) GAN. (e) CNN-MSE. (f) CNN-VGG. (g) WGAN. (h) WGAN-MSE. (i) WGAN-VGG.

TABLE II
QUANTITATIVE RESULTS ASSOCIATED WITH DIFFERENT NETWORK OUTPUTS FOR FIGS. 5 AND 7

	Fig. 5		Fig. 7	
	PSNR	SSIM	PSNR	SSIM
LDCT	19.7904	0.7496	18.4519	0.6471
CNN-MSE	24.4894	0.7966	23.2649	0.7022
WGAN-MSE	24.0637	0.8090	22.7255	0.7122
CNN-VGG	23.2322	0.7926	22.0950	0.6972
WGAN-VGG	23.3942	0.7923	22.1620	0.6949
WGAN	22.0168	0.7745	20.9051	0.6759
‘1 GAN	21.8676	0.7581	21.0042	0.6632
DictRecon	24.2516	0.8148	24.0992	0.7631

to the per-pixel loss, it is not surprising that CNN-MSE, which was trained to minimize MSE loss, outperformed the networks trained to minimize other feature-based loss. It is worth noting that these quantitative results are in decent agreement with Fig. 4, in which CNN-MSE has the smallest MSE loss and WGAN has the largest. The reason why WGAN ranks the worst in PSNR and SSIM is because it does not include either MSE or VGG regularization. DictRecon achieves the best SSIM and a high PSNR. However, it has the problem of image blurring and leads to blocky and waxy artifacts in the resultant images. This indicates that PSNR and SSIM may not be sufficient in evaluating image quality.

TABLE III
STATISTICAL PROPERTIES OF THE BLUE RECTANGLE AREAS IN FIGS. 5 AND 7. THE VALUES ARE IN HOUNSFIELD UNIT (HU)

	Fig. 5		Fig. 7	
	Mean	SD	Mean	SD
NDCT	9	36	118	38
LDCT	11	74	118	66
CNN-MSE	12	18	120	15
WGAN-MSE	9	28	115	25
CNN-VGG	4	30	104	28
WGAN-VGG	9	31	111	29
WGAN	23	37	135	33
GAN	8	35	110	32
DictRecon	4	11	111	13

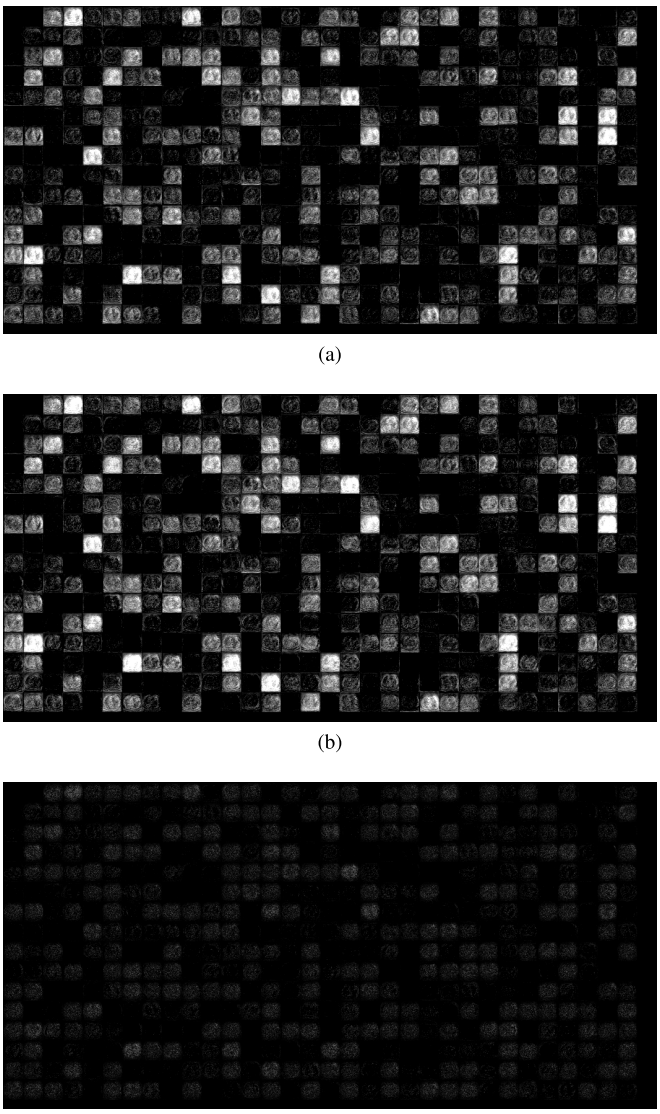
In the reviewing process, we found two papers using similar network structures. Wolterink *et al.* [38] trained three networks, i.e. GAN, CNN-MSE, and GAN-MSE for cardiac CT denoising. Their quantitative PSNR results are consistent with our counterpart results. And Yu *et al.* [39] used GAN-VGG to handle the de-alising problem for fast CS-MRI. Their results are also consistent with ours. Interestingly, despite the high PSNRs obtained by MSE-based networks, the authors in the two papers all claim that GAN and VGG loss based networks have better image quality and diagnostic information.

To gain more insight into the output images from different approaches, we inspect the statistical properties by calculating the mean CT numbers (Hounsfield Units) and standard deviations (SDs) of two flat regions in Figs. 5 and 7 (marked by the blue rectangles). In an ideal scenario, a noise reduction algorithm should achieve mean and SD to the gold standard as close as possible. In our experiments, the NDCT FBP images were used as gold standard because they have the best image quality in this dataset. As shown in Table III, Both CNN-MSE and DictRecon produced much smaller SDs compared to NDCT, which indicates they over-smoothed the images and supports our visual observation. On the contrary, WGAN produced the closest SDs yet smaller mean values, which means it can reduce noise to the same level as NDCT but it compromised the information content. On the other hand, the proposed WGAN-VGG has outperformed CNN-VGG, WGAN-MSE and other selected methods in terms of mean CT numbers, SDs, and most importantly visual impression.

In addition, we performed a blind reader study on 10 groups of images. Each group contains the same image slice but processed by different methods. NDCT and LDCT images are also included for reference, which are the only two labeled images in each group. Two radiologists were asked to independently score each image in terms of noise suppression and artifact reduction on a five-point scale (1 = unacceptable and 5 = excellent), except for the NDCT and LDCT images, which are the references. In addition, they were asked to give an overall image quality score for all the images. The mean and standard deviation values of the scores from the two radiologists were then obtained as the final evaluation results, which are shown in Table. IV. It can be seen that CNN-MSE

TABLE IV
SUBJECTIVE QUALITY SCORES (MEAN±SD) FOR DIFFERENT ALGORITHMS

	NDCT	LDCT	CNN-MSE	CNN-VGG	WGAN-MSE	WGAN-VGG	WGAN	GAN	DictRecon
Noise Suppression	-	-	4.35 ± 0.24	3.10 ± 0.23	3.55 ± 0.25	3.20 ± 0.25	2.90 ± 0.26	3.00 ± 0.21	$\mathbf{4.65 \pm 0.20}$
Artifact Reduction	-	-	1.70 ± 0.28	2.85 ± 0.32	3.05 ± 0.27	$\mathbf{3.45 \pm 0.25}$	2.90 ± 0.28	3.05 ± 0.27	2.05 ± 0.27
Overall Quality	3.95 ± 0.20	1.35 ± 0.16	2.15 ± 0.25	3.05 ± 0.20	3.30 ± 0.21	$\mathbf{3.70 \pm 0.15}$	3.05 ± 0.22	3.10 ± 0.21	2.05 ± 0.36



(a) (b) (c)

Fig. 9. VGG feature maps of full dose and quarter dose images in Fig. 5 and their absolute difference. (a) VGG Map of Full Dose Image. (b) VGG Map of Quarter Dose Image. (c) Absolute Difference.

and DictRecon give the best noise suppression scores while the proposed WGAN-VGG outperforms the other methods for artifact reduction and overall quality improvement. Also, *-VGG networks provide higher scores than *-MSE networks in terms of artifact reduction and overall quality but lower scores for noise suppression. This indicates that MSE loss based networks are good at noise suppression at a loss of image details, resulting in an image quality degradation for diagnosis. Meanwhile, the networks using WGAN give better overall image quality than the networks using CNN, which supports the use of WGAN for CT image denoising.

F. VGG Feature Extractor

Since VGG network is trained on natural images, it may cause concerns on how well it performs on CT image feature extraction. Thus, we displayed two feature maps of normal dose and quarter dose images and their absolute difference in Fig. 9. The feature map contains 512 small images of size 32×32. We organize these small images into a 32×16 array. Each small image emphasizes a feature of the original CT image, i.e. boundaries, edges, or whole structures. Thus, we believe VGG network can also serve a good feature extractor for CT images.

IV. DISCUSSIONS AND CONCLUSION

The most important motivation for this paper is to approach the gold standard NDCT images as much as possible. As described above, the feasibility and merits of GAN has been investigated for this purpose with the Wasserstein distance and the VGG loss. The difference between using the MSE and VGG losses is rather significant. Despite the fact that networks with MSE would offer higher values for traditional figures of merit, VGG loss based networks seem desirable for better visual image quality with more details and less artifacts.

The experimental results have demonstrated that using WGAN helps improve image quality and statistical properties. Comparing the images of CNN-MSE and WGAN-MSE, we can see that the WGAN framework helped to avoid over-smoothing effect typically suffered by MSE based image generators. Although CNN-VGG and WGAN-VGG visually share a similar result, the quantitative analysis shows WGAN-VGG enjoys higher PSNRs and more faithful statistical properties of denoised images relative to those of NDCT images. However, using WGAN/GAN alone reduced noise but at the expense of losing critical features. The resultant images do not show a strong noise reduction. Quantitatively, the associated PSNR and SSIM increased modestly compared to LDCT but they are much lower than what the other networks produced. Theoretically, WGAN/GAN network is based on generative model and may generate images that look naturally yet cause a severe distortion for medical diagnostics. This is why an additive loss function such as MSE and VGG loss should be added to guarantee the image content remains the same.

It should be noted that the experimental data contain only one noise setting. Networks should be re-trained or re-tuned for different data to adapt for different noise properties. Especially, networks with WGAN are trying to minimize the

distance between two probability distributions. Thus, their trained parameters have to be adjusted for new datasets. Meanwhile, since the loss function of WGAN-VGG is a mixture of feature domain distance and the GAN adversarial loss, they should be carefully balanced for different dataset to reduce the amount of image content alternation.

The denoising network is a typical end-to-end operation, in which the input is a LDCT image while the target is a NDCT image. Although we have generated images visually similar to NDCT counterparts in the WGAN-VGG network, we recognize that these generated images are still not as good as NDCT images. Moreover, noise still exists in NDCT images. Thus, it is possible that VGG network has captured these noise features and kept them in the denoised images. This could be a common problem for all the denoising networks. How to outperform the so-called gold standard NDCT images is an interesting open question. Moreover, image post-denoising methods also suffer from the information loss during the FBP reconstruction process. This phenomena is observed in the comparison with DictRecon result. A better way to incorporate the strong fitting capability of neural network and the data completeness of CT data is to design a network that maps directly from raw projection to the final CT images, which could be a next step of our work.

In conclusion, we have proposed a contemporary deep neural network that uses a WGAN framework with perceptual loss function for LDCT image denoising. Instead of focusing on the design of a complex network structure, we have dedicated our effort to combine synergistic loss functions that guide the denoising process so that the resultant denoised results are as close to the gold standard as possible. Our experiment results with real clinical images have shown that the proposed WGAN-VGG network can effectively solve the well-known over-smoothing problem and generate images with reduced noise and increased contrast for improved lesion detection. In the future, we plan to incorporate the WGAN-VGG network with more complicated generators such as the networks reported in [26] and [27] and extend these networks for image reconstruction from raw data by making a neural network counterpart of the FBP process.

V. Acknowledgment

The authors would also like to thank NVIDIA Corporation for the donation of the Titan Xp GPU used for this research.

References

[1] D. J. Brenner and E. J. Hall, "Computed tomography—An increasing source of radiation exposure," *New England J. Med.*, vol. 357, no. 22, pp. 2277–2284, Nov. 2007.

[2] A. B. de González and S. Darby, "Risk of cancer from diagnostic X-rays: Estimates for the UK and 14 other countries," *Lancet*, vol. 363, no. 9406, pp. 345–351, Jan. 2004.

[3] J. Wang, H. Lu, T. Li, and Z. Liang, "Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters," *Proc. SPIE*, vol. 5747, pp. 2058–2067, Apr. 2005.

[4] J. Wang, T. Li, H. Lu, and Z. Liang, "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography," *IEEE Trans. Med. Imag.*, vol. 25, no. 10, pp. 1272–1283, Oct. 2006.

[5] A. Manduca *et al.*, "Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT," *Med. Phys.*, vol. 36, no. 11, pp. 4911–4919, 2009.

[6] M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in X-ray CT," *Phys. Med.*, vol. 28, no. 2, pp. 94–108, Apr. 2012.

[7] A. K. Hara, R. G. Paden, A. C. Silva, J. L. Kujak, H. J. Lawder, and W. Pavlicek, "Iterative reconstruction technique for reducing body radiation dose at CT: Feasibility study," *Amer. J. Roentgenol.*, vol. 193, no. 3, pp. 764–771, Sep. 2009.

[8] J. Ma *et al.*, "Low-dose computed tomography image restoration using previous normal-dose scan," *Med. Phys.*, vol. 38, no. 10, pp. 5713–5731, 2011.

[9] Y. Chen *et al.*, "Improving abdomen tumor low-dose CT images using a fast dictionary learning based processing," *Phys. Med. Biol.*, vol. 58, no. 16, p. 5803, Aug. 2013.

[10] P. F. Feruglio, C. Vinegoni, J. Gros, A. Sbarbati, and R. Weissleder, "Block matching 3D random noise filtering for absorption optical projection tomography," *Phys. Med. Biol.*, vol. 55, no. 18, p. 5401, Sep. 2010.

[11] B. De Man and S. Basu, "Distance-driven projection and backprojection in three dimensions," *Phys. Med. Biol.*, vol. 49, no. 11, p. 2463, 2004.

[12] R. M. Lewitt, "Multidimensional digital image representations using generalized Kaiser–Bessel window functions," *J. Opt. Soc. Amer. A*, vol. 7, no. 10, pp. 1834–1846, Oct. 1990.

[13] B. R. Whiting, P. Massoumzadeh, O. A. Earl, J. A. O'Sullivan, D. L. Snyder, and J. F. Williamson, "Properties of preprocessed sinogram data in X-ray computed tomography," *Med. Phys.*, vol. 33, no. 9, pp. 3290–3303, Sep. 2006.

[14] I. A. Elbakri and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," *IEEE Trans. Med. Imag.*, vol. 21, no. 2, pp. 89–99, Feb. 2002.

[15] S. Ramani and J. A. Fessler, "A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 677–688, Mar. 2012.

[16] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Phys. Med. Biol.*, vol. 53, no. 17, p. 4777, Sep. 2008.

[17] Y. Liu, J. Ma, Y. Fan, and Z. Liang, "Adaptive-weighted total variation minimization for sparse data toward low-dose X-ray computed tomography image reconstruction," *Phys. Med. Biol.*, vol. 57, no. 23, p. 7923, 2012.

[18] Z. Tian, X. Jia, K. Yuan, T. Pan, and S. B. Jiang, "Low-dose CT reconstruction via edge-preserving total variation regularization," *Phys. Med. Biol.*, vol. 56, no. 18, p. 5949, Nov. 2011.

[19] 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. Imaging*, vol. 31, no. 9, pp. 1682–1697, Sep. 2012.

[20] Y. Zhang, X. Mou, G. Wang, and H. Yu, "Tensor-based dictionary learning for spectral CT reconstruction," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 142–154, Jan. 2017.

[21] D. Kang *et al.*, "Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm," *Proc. SPIE*, vol. 8669, p. 86692G, Mar. 2013.

[22] G. Wang, M. Kalra, and C. G. Orton, "Machine learning will transform radiology significantly within the next 5 years," *Med. Phys.*, vol. 44, no. 6, pp. 2041–2044, 2017.

[23] G. Wang, "A perspective on deep imaging," *IEEE Access*, vol. 4, pp. 8914–8924, 2016.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[25] H. Chen *et al.* (2016). *Low-Dose CT Denoising With Convolutional Neural Network*. [Online]. Available: https://www.Figures:1610.00321

[26] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Image Process.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017.

[27] E. Kang, J. Min, and J. C. Ye. (2016). "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction." [Online]. Available: https://arxiv.org/abs/1610.09736

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[29] J. Johnson, A. Alahi, and L. Fei-Fei. (2016). "Perceptual losses for real-time style transfer and super-resolution." [Online]. Available: https://arxiv.org/abs/1603.08155

[30] C. Ledig *et al.* (2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802
[31] M. Arjovsky, S. Chintala, and L. Bottou. (2017). "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875
[32] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[33] I. Goodfellow. (2017). "NIPS 2016 tutorial: Generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1701.00160
[34] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. (2016). "Neural photo editing with introspective adversarial networks." [Online]. Available: https://arxiv.org/abs/1609.07093
[35] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.
[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004
[37] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen. (2016). "Medical image synthesis with context-aware generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1612.05362
[38] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017.
[39] S. Yu *et al.* (2017). "Deep de-aliasing for fast compressive sensing MRI." [Online]. Available: https://arxiv.org/abs/1705.07137
[40] M. Arjovsky and L. Bottou. (2017). "Towards principled methods for training generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1701.04862
[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. (2017). "Improved training of wasserstein GANs." [Online]. Available: https://arxiv.org/abs/1704.00028
[42] M. Nixon and A. S. Aguado, *Feature Extraction & Image Process*, 2nd ed. San Francisco, CA, USA: Academic, 2008.
[43] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[44] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers Robot. AI*, vol. 2, p. 36, Jan. 2016.
[45] AAPM. (2017). *Low Dose CT Grand Challenge*. [Online]. Available: http://www.aapm.org/GrandChallenge/LowDoseCT/#
[46] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[47] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467
[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.