SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network

Meng Li, *Member, IEEE*, William Hsu, *Member, IEEE*, Xiaodong Xie, Jason Cong, *Fellow, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— Computed tomography (CT) is a widely used screening and diagnostic tool that allows clinicians to obtain a high-resolution, volumetric image of internal structures in a non-invasive manner. Increasingly, efforts have been made to improve the image quality of low-dose CT (LDCT) to reduce the cumulative radiation exposure of patients undergoing routine screening exams. The resurgence of deep learning has yielded a new approach for noise reduction by training a deep multi-layer convolutional neural networks (CNN) to map the low-dose to normal-dose CT images. However, CNN-based methods heavily rely on convolutional kernels, which use fixed-size filters to process one local neighborhood within the receptive field at a time. As a result, they are not efficient at retrieving structural information across large regions. In this paper, we propose a novel 3D self-attention convolutional neural network for the LDCT denoising problem. Our 3D self-attention module leverages the 3D volume of CT images to capture a wide range of spatial information both within CT slices and between CT slices. With the help of the 3D self-attention module, CNNs are able to leverage pixels with stronger relationships regardless of their distance and achieve better denoising results. In addition, we propose a self-supervised learning scheme to train a domain-specific autoencoder as the perceptual loss function. We combine these two methods and demonstrate their effectiveness on both CNN-based neural networks and WGAN-based neural networks with comprehensive experiments. Tested on the AAPM-Mayo Clinic Low Dose CT Grand Challenge data set, our experiments demonstrate that self-attention (SA) module and autoencoder (AE) perceptual loss function can efficiently enhance traditional CNNs and can achieve comparable or better results than the state-of-the-art methods.**

***Index Terms*— Low-dose CT, denoising, self-attention, autoencoder, perceptual loss.**

Manuscript received December 5, 2019; accepted January 6, 2020. Date of publication January 21, 2020; date of current version June 30, 2020. This work was supported in part by the National Key Research and Development Program of China under Contract 2016YFB0402001, in part by the Beijing Major Science and Technology Project under Grant Z191100010618003, in part by the National Natural Science Foundation of China (NSFC) under Grant 61520106004, and in part by the PKU-UCLA Joint Research Institute. *(Corresponding author: Meng Li.)*

Meng Li, Xiaodong Xie, and Wen Gao are with the Department of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: mmli@pku.edu.cn; donxie@pku.edu.cn; wgao@pku.edu.cn).

William Hsu is with the Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA 90024 USA (e-mail: whsu@mednet.ucla.edu).

Jason Cong is with the Department of Computer Science, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: cong@cs.ucla.edu).

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

Digital Object Identifier 10.1109/TMI.2020.2968472

I. Introduction

COMPUTED tomography (CT) is a versatile, high resolution imaging modality that is increasingly used in screening and diagnostic applications such as the detection of pulmonary nodules. Its ability to resolve small objects, such as nodules that are < 30 mm in size, makes the modality indispensable in performing tasks like lung cancer screening. However, as patients are imaged using CT over time, the cumulative exposure to radiation brings potential health risks [1], [2]. Given these risks, efforts have been made on reducing the radiation dose that a patient is exposed to in a given CT study. In general, a lower radiation dose can be achieved by reducing the tube current or shortening the exposure time of the X-ray tube. Nevertheless, reducing x-ray exposure is associated with an increase in noise and artifacts of reconstructed images. Efforts have been made to improve the image quality of low-dose CT (LDCT).

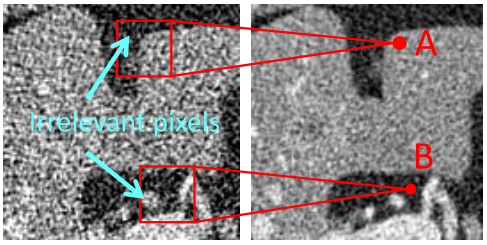
There are three categories of algorithms in general: a) sinogram domain filtration, b) iterative reconstruction, and c) image post-processing. Sinogram filtering directly smooths raw data before reconstruction (i.e., filtered back projection, FBP). In [3], Manduca *et al.* proposed bilateral filtering incorporating a CT noise model that achieved a significantly better noise-resolution trade-off than comparison commercial reconstruction kernels. Some other typical methods include structural adaptive filtering [4] and penalized weighted least-squares algorithms [5]. Li *et al.* proposed a statistical model of sinogram data, developing a penalized likelihood method to suppress quantum noise [6]. Sinogram filtering algorithms are often restricted in practice due to the difficulty of accessing projection data. Iterative reconstruction (IR) algorithms estimate the denoised images incrementally using prior information from the image domain. Several image priors such as total variation (TV) [7]–[10], non local means [11]–[13], dictionary learning [14], [15], and other techniques [16]–[18] have been formulated. Although these iterative reconstruction

algorithms have greatly improved denoising performance, they incur a high computational cost on iterative projection/backprojection steps, which increases the amount of time required to generate the reconstructed volumes.

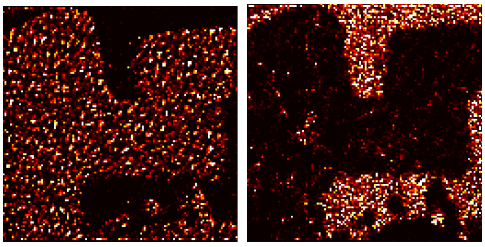
Recently, researchers have explored image denoising-based techniques by post-processing the reconstructed CT images instead of directly processing the raw projection data. Initial studies tried classical image-processing algorithms like non-local means filtering [19]–[21], dictionary-learning-based methods [22], [23], block-matching algorithms [24]–[26], and diffusion filters [27], which are more computationally efficient than IR methods. However, the noise in reconstructed low-dose CT images often has a non-uniform distribution, which is too complex to be dealt with by these methods.

The resurgence of deep neural networks has yielded a new approach for noise reduction, which learns a nonlinear transformation function from low-dose and normal-dose CT image pairs. Convolutional neural networks (CNN) are a class of hierarchical multi-layer deep neural networks with non-linear and convolutional filters, and have been demonstrated to be highly effective in a variety of tasks such as image classification, denoising, and super-resolution [28]–[31]. Early research in LDCT denoising focused on CNN architecture optimization and adaptation, such as RED-CNN [32] and Wavelet networks [33]. However, the MSE-loss used in previous methods tended to generate overly smoothed images. To solve this problem, Johnson *et al.* [34] proposed the perceptual-loss based on a pre-trained VGG model [35]. Subsequently, a generative adversarial network (GAN) [36], [37] was introduced to overcome the limitations of voxel-wise regression in noise reduction [38]–[41], where a generative network, G, was trained to produce realistic images while a discriminator network, D, was trained to classify whether an image is real or generated.

Despite the progress on denoising LDCT images, CNN-based methods heavily rely on convolutional kernels to model the correlated structural information. Convolutional kernels are building blocks that process one *local* neighborhood within a *fixed-size* receptive field, usually 3x3 or 5x5 in size, at a time. This localized and fixed-size convolutional kernel is inefficient at processing structural information across large regions, or namely *long-range dependencies* [42], [43]. Two reasons are provided to explain this inefficiency. First, as Fig. 1a shows, fixed-size filtering unavoidably involves irrelevant pixels for the current response, especially for the edges or regions with complex structural patterns. Second, although traditional convolution is also able to obtain a larger receptive field by passing through multiple layers, the volume of irrelevant pixels introduced into the scope is enlarged as well and results in loses in the computational and statistical efficiency. In addition, the training algorithms may have trouble in coordinating the carried dependencies across multiple layers, resulting in inefficient weight learning. A self-attention model, however, solves these problems by establishing relationships between the local response and all other pixels *within* one layer to guide the convolutional filtering. Fig. 1b and Fig. 1c visualize the long-range dependencies of pixel A and B that are generated from our self-attention



(a) Local receptive field



(b) Pixel A's related pixels (c) Pixel B's related pixels

Fig. 1. Fixed-sized local receptive field and long-range dependencies. (a) Filtering pixels equally in a fixed-size local receptive field can introduce irrelevant pixels that are uninformative to compute the denoising output, especially for edges and regions with complex structural patterns. (b) The long-range dependencies of pixel A in the image. (c) The long-range dependencies of pixel B. The highlight color means stronger relationships and the dark color means weaker relationships. (b) and (c) show attention feature maps from the proposed self-attention module when given an input CT image.

module when given an input CT image, which will be further discussed in Section II. From these figures, we can see pixel A is more relevant to pixels in the organs and B is more relevant to background pixels. Related work in natural images [43], [44] showed that convolution-based methods are good at synthesizing the classes of image with more texture pattern and few structural constraints (e.g. grassland, sky, and seas) but poor at synthesizing images with geometric or structural patterns (e.g. patterned skins of animals). CT images are usually a mixture of geometric shapes with textures and non-uniformly distributed noise, which adds to the complexity of the denoising problem. As a result, CNN-based methods are inherently less efficient in modeling various structural information in CT images.

In traditional image denoising, the idea of non-local means [45] was proposed to enable distant pixels to contribute to the filtered response at a location based on patch similarity. This idea was widely used in applications such as image denoising [46], texture synthesis [47], super-resolution [48], and inpainting [49]. As studied in [42], self-attention can be regarded as a type of non-local operation with trainable parameters. The attention mechanism was first proposed in [50], where it used different positions of a single sequence to compute a representation of the sequence. The self-attention is also widely used in computer vision applications, such as image synthesis [43], action classification [51] and scene segmentation [52] by capturing rich contextual dependencies.

However, self-attention in previous work was primarily used to process 2D images. Capturing long-range dependencies between slices in spatial 3D CT images remains underexplored. To better exploit the 3D spatial information in

CT images, we introduce a 3D self-attention model, combining two types of attention modules to capture global relationships in the spatial domain. We refer to these modules as *plane attention* and *depth attention*, which deal with long-range dependencies within a CT slice and between CT slices, respectively. The proposed 3D self-attention neural network enables convolution layers to capture long-range dependencies in 3D CT images and can further improve the performance of image denoising.

In addition, we propose a self-supervised learning scheme to train a better perceptual loss function. Perceptual loss [34] was first proposed to replace the MSE method for better image feature comparison, which uses VGG model as the feature extractor. Yang's work [53] extended this VGG-loss method to train a Wasserstein GAN [37]. However, we believe for CT images, VGG feature extractors tend to lose important domain-specific details because it is originally trained for classifying natural images. Therefore, VGG-loss is not the best perceptual loss function for LDCT denoising. The major challenge of training a domain-specific perceptual network is that CT data set is often unlabeled, which makes it impossible to apply supervised learning to either train a new neural network from scratch or transfer an existing deep learning model. To solve this problem, we propose a self-supervised learning scheme to train a perceptual loss neural network specifically for CT images. We design an autoencoder neural network [54], [55] to learn an efficient manifold from normal-dose CT (NDCT) data and then use the encoding features to compute a loss function.

The main contribution of this paper are as follows:

- We propose a novel 3D self-attention convolutional neural network for LDCT image denoising. By explicitly using two types of attention modules, we efficiently capture long-range dependencies in 3D LDCT images and demonstrate meaningful improvement on the denoising results.
- We propose an autoencoder neural network to train a perceptual loss function that can better extract and evaluate CT features. Using the domain-specific perceptual loss, we can effectively expand the network representational ability and further improve denoising performance.
- We demonstrate the effectiveness of the self-attention module and autoencoder with extensive experiments, and also show that the proposed methods improve denoising performance in both CNN-based and GAN-based neural networks. In addition, we perform quantitative and qualitative comparisons between other state-of-the-art methods.

The remainder of the paper is organized as follows: In section II, we first describe the CT denoising problem and propose our self-attention method and autoencoder perceptual network, introducing the overall network structure. In section III, we evaluate our proposed network, performing ablation studies and comparing our model's results against current state-of-the-art methods, including RED-CNN [32], GAN [38], WGAN-VGG [53] and CPCE-3D [41]. In section IV, we discuss the implications, limitations, future work, and key contributions.

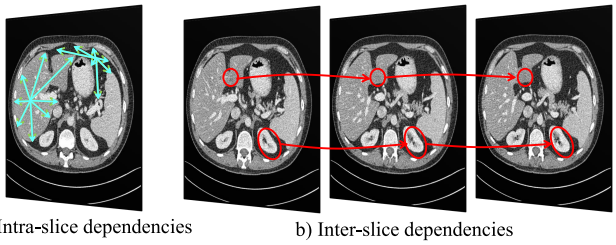


Fig. 2. Two types of long-range dependencies in 3D CT images. (a) Intra-slice dependencies are relationships of pixels *within* a CT slice. For example, pixels from the same organ (or black background) are highly related. (b) Inter-slice dependencies are relationships of pixels *between* CT slices. For example, pixels of the same organ (or the same lesion) among different CT slices are highly related.

II. Method

The LDCT denoising problem can be formulated as a noise reduction model in the image spatial domain. Let $I_{LD} \in \mathbb{R}^{D\times H\times W}$ denotes a LDCT image, where D, H, W represent depth, height, and width of the image, respectively. $I_{ND} \in \mathbb{R}^{D\times H\times W}$ denotes a NDCT image. The goal of denoising is to seek a generator function g that maps a LDCT image to a NDCT image:

$$g(I_{LD}) = I_{ND} \tag{1}$$

Instead of directly removing the noise from the LDCT image, we adopt a CNN model to extract key features from LDCT (I_{LD}) and synthesize a new image that is close to the NDCT (I_{ND}). In our scheme, we incorporate a 3D self-attention (SA) module into a 3D convolutional neural network to generate the NDCT output. We call this model a SACNN generator. In order to optimize the SACNN generator $g(\cdot)$, we use a joint optimization objective function which consists of a discriminator CNN and a perceptual loss network. We use a WGAN [56] training scheme to optimize this joint loss function.

We introduce the 3D self-attention CNN, self-supervised autoencoder perceptual loss, and overall architecture in the following sections.

A. 3D Self-Attention Convolutional Neural Network (SACNN)

A self-attention module computes the correlation matrix that represents spatial dependencies between any two positions within the input feature maps. Each position is calculated and updated by the weighted sum of all other positions. The weight values are decided by learning dependencies between the two positions. Therefore, any two positions with similar features or strong dependencies will be represented in the correlation matrix and mutually contribute to the final response regardless of their distance on the input image or feature maps.

We identify and differentiate two types of dependencies in CT images: *intra-slice dependency* and *inter-slice dependency*. As shown in Fig. 2, intra-slice dependencies are relationships of pixels *within* a CT slice, such as pixels of the same organ. Inter-slice dependencies are relationships of pixels *between* CT slices, such as the same organ among different CT slices.

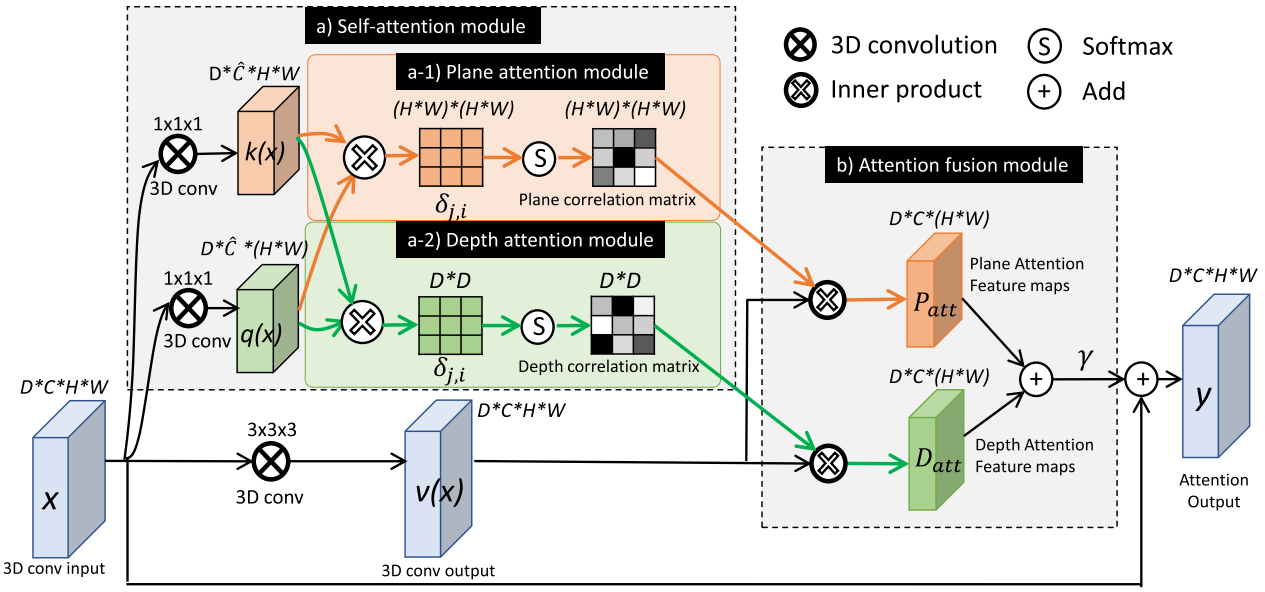


Fig. 3. The proposed 3D self-attention convolutional neural network.

We propose a 3D self-attention module, including *plane attention module* to deal with the intra-slice dependencies and *depth attention module* to deal with the inter-slice dependencies. We describe them in detail in the following paragraphs.

We define the feature maps in the current hidden layer as $x \in \mathbb{R}^{D\times C\times H\times W}$, where $\langle D\rangle$ denotes depth (the CT slices number), $\langle C\rangle$ denotes channel number (the number of feature maps), $\langle H, W\rangle$ denotes height and width of a CT slice or a feature map, as is shown in Fig. 3. We first transform the input feature maps X into two feature spaces $q(x) \in \mathbb{R}^{\hat{C}\times D\times H\times W}$ and $k(x) \in \mathbb{R}^{\hat{C}\times D\times H\times W}$ by applying a $1\times 1\times 1$ convolution. This transformation serves as an feature encoding and channel pooling for original input, following the approach in [50], [57], so that $\hat{C}$ is a hyper-parameter determined empirically and $\hat{C} \le C$.

1) Plane Attention Module: For intra-slice dependencies, a plane attention module is used to compute the relationships of each pixel to all the other pixels within a CT slice. We first flatten the 2D array into 1D vector by applying an array reshape operation from $\hat{C}\times D\times H\times W$ to $\hat{C}\times D\times N$, where $N = H\times W$ is the total number of pixels in a feature map. Then we perform an inner product of the two vectors $q(x_i)$ and $k(x_j)$ on every channel to get the relationship between every two positions.

$$\delta_{j,i} = q(x_i)^T k(x_j) \tag{2}$$

where $\delta_{j,i} \in \mathbb{R}^{\hat{C}\times D\times N\times N}$. $\delta_{j,i}$ indicates the relationships or similarity of i^{th} location to j^{th} location on feature maps, where large values mean strong relationships and small values mean weak relationships. Then we use the softmax function to normalize δ.

$$r_{j,i} = Softmax(\delta_{j,i}) = \frac{e^{\delta_{j,i}}}{\sum_{i=1}^{N} e^{\delta_{j,i}}} \tag{3}$$

Finally, we apply the correlation matrix to the convolution feature maps to calculate the response of the plane attention, which can be formulated as:

$$P_{att} = \sum_{i=1}^{N} v(x_i)\times r_{j,i} \tag{4}$$

where $v(x) \in \mathbb{R}^{C\times D\times H\times W}$ is the backbone convolutional layer with a $3\times 3\times 3$ kernel. The structure of plane attention is depicted in Fig. 3 (a-1).

2) Depth Attention Module: In order to relate global features from other CT slices, we build a depth attention module to explicitly model the inter-slice dependencies. The structure of depth attention module is shown in Fig. 3 (a-2). Different from the plane attention, depth attention computes the similarity matrix along the depth dimension. Specifically, we first transpose the $q(x)$ and $k(x)$ from $D\times C\times H\times W$ to $(H\times W)\times D\times \hat{C}$, with which we apply the inner product like (2). The result is a depth correlation matrix $\delta_{j,i}$ with dimension $\mathbb{R}^{D\times D}$ after softmax normalization. The $D\times D$ matrix represents the relationships of pixels to each other on the depth dimension.

3) Attention Fusion Module: In order to fuse the plane and depth information from two attention modules, we aggregate the plane attention P_{att} and depth attention D_{att} at the end of the self-attention module, shown in Fig. 3 (b). We multiply the output of the attention layer by a trainable scalar parameter, adding the result back to the input feature map as a residual block.

$$Y = \gamma\,(P_{att} + D_{att}) + X \tag{5}$$

where the parameter γ is a scaling factor and is trainable by back propagation algorithm. γ is initialized as 0 so the network will start with classical CNNs. With the training process goes on, it learns more spatial information from the training data, where γ will gradually apply the self-attention features. This is easier for the network to train and achieve a good convergence.

4) SACNN Generator: As illustrated in Fig. 4, our SACNN generator is composed of multiple layers of the self-attention

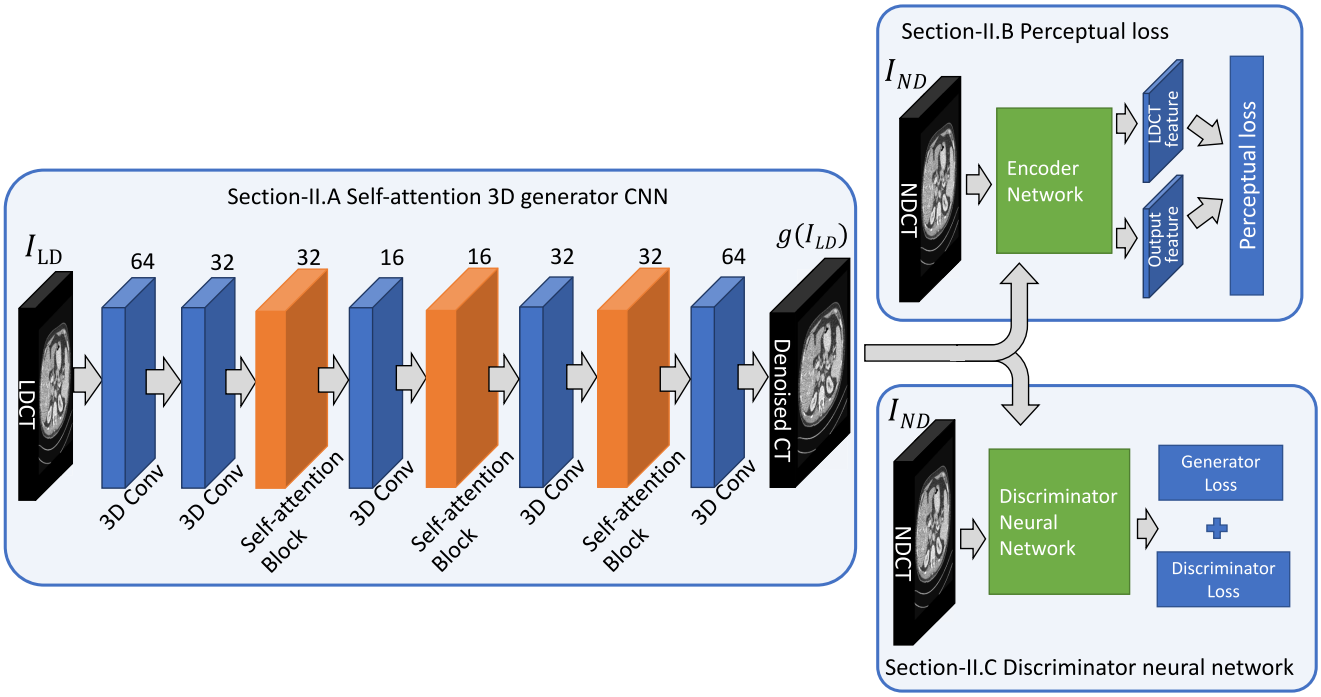


Fig. 4. Overall network architecture of proposed method.

block. The generator network consists of 8 convolutional layers with 64, 32, 32, 16, 16, 32, 32, and 64 filters. For the input layer, we compose 3 input slices to the 3D extension and apply to later convolutional layers. For each hidden layers and output layer, the feature maps are 4-D tensor in terms of $[channel, D, H, W]$.

Each layer is followed by a ReLU function except for the input of the self-attention block. A $3\times3\times3$ filter is used for all convolutional layers. We employ three self-attention block in the generator SACNN network at the third, fifth and seventh layer as shown in Fig. 4. This cascade connection structure enables the network to reduce the noise information gradually and then restore the clean image.

B. Self-Supervised Learning for Perceptual Loss

1) Perceptual Loss: Perceptual loss measures the similarity between images by pre-processing the input image through a convolutional neural network. Image denoising neural networks that use perceptual loss are demonstrated to be more robust to many potential issues like over-smoothing and distortion compared to mean-squared error (MSE) loss [34], [53]. In previous work [41], [53], several convolution layers from VGG-19 were used, which is a pre-trained neural network on a natural image dataset [58], to compute the perceptual loss and achieve an impressive result. We refer it as VGG-loss, which is formulated as the following,

$$L_{VGG}(g) = \mathbb{E}_{(I_{LD},I_{ND})}[\frac{||VGG(g(I_{LD})) - VGG(I_{ND})||_F^2}{DHW}] \tag{6}$$

where D, H, W represents depth, height, and width of the CT image.

2) Autoencoder Perceptual Loss: A potential problem of using VGG-loss is that the original VGG network was trained for image classification using natural images. VGG-based

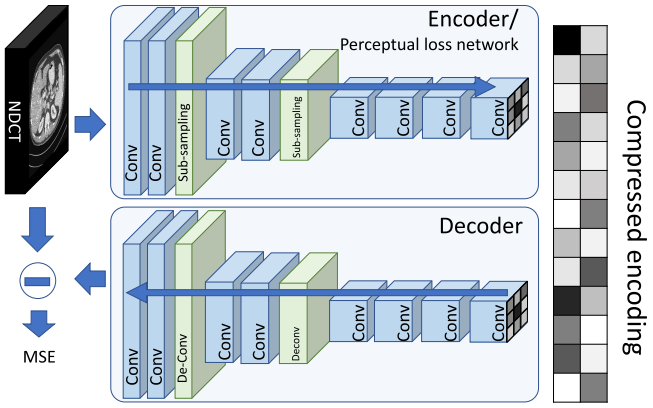


Fig. 5. The proposed autoencoder network.

feature extractors may generate features that are not relevant to CT image denoising. However, retraining VGG for CT data sets is challenging because there is not a collection of labeled medical images that can match the size of ImageNet. To solve this problem, we adopt a self-supervised learning scheme to train an autoencoder network. The target of autoencoder neural network is to learn to extract a compressed encoding from the input CT data, and then reconstruct it into an image that is closely similar to the original input. Our autoencoder is composed of two sub neural networks - *Encoder network* and *Decoder network*. In our perceptual loss network in Fig. 4, we only use the encoder network and its output compressed encoding as the perceptual loss. The decoder network is only associative for training the encoder network.

As shown in Fig. 5, the Encoder network consists of 8 convolution layers with 64, 64, 128, 128, 256, 256, 256, and 256 filters, respectively. We insert two max pooling layer after the second and fourth layer, to compress the features to a lower dimension. The kernel size and stride number of max pooling is 2. Note that convolutional layers are followed

with a ReLU activation function. We adopt a 3×3 filter size in this design and a stride of 1. The decoder network is completely symmetrical with the encoder network. For the max pooling layer, we adopt a transposed convolution layer to upscale the image 2 times along each dimension. The MSE objective optimization procedure uses gradient descent algorithm to enforce the autoencoder to learn a dimension transformation to filter out noises.

After training this network, we adopt the output of encoder network as the extracted feature and use it in the proposed perceptual loss function:

$$L_{AE}(g) = \mathbb{E}_{(I_{LD}, I_{ND})}[\frac{||\phi(g(I_{LD})) - \phi(I_{ND})||_F^2}{DHW}] \quad (7)$$

where ϕ is the pre-trained Encoder network.

C. Network Structure

Inspired by work [38], [56], we adopt the WGAN framework, whose optimization objective can be formulated as the following equation,

$$\min_g \max_d L_{WGAN}(g, d) + \eta L_{AE}(g) \quad (8)$$

where $L_{WGAN}(g, d)$ is WGAN's optimization objective, $L_{AE}(g)$ is perceptual loss, g is the generator network, d is the discriminator network, η is a weighted parameter to control the trade-off between WGAN loss function and the autoencoder perceptual loss. $L_{WGAN}(g, d)$ is an addition of traditional Wasserstein distance and the gradient penalty for regularization, which is expressed as

$$\begin{aligned} \min_g \max_d L_{WGAN}(g, d) = (&-\mathbb{E}_{I_{ND}}[d(I_{ND})] \\ &+\mathbb{E}_{I_{LD}}[d(g(I_{LD}))]) \\ &+\lambda \mathbb{E}_{\hat{I}}[(||\nabla_{\hat{I}} d(\hat{I})||_2 - 1)^2] \end{aligned} \quad (9)$$

where $\mathbb{E}_a[b]$ denotes the expectation of b, as a function of a, λ is a weighting parameter, the $\hat{I}$ represents the generated and real images in a uniformly sampling from an interval of [0, 1], and ∇ denotes the gradient. The generator network and discriminator network are trained in an alternating fashion by minimizing the hinge version of the adversarial loss.

The overall network architecture of our method is depicted in Fig. 4, which consists of three components. We use 3D self-attention CNN network (SACNN) as the generator network (g), which has been introduced in section II-A. We train the autoencoder network with self-supervised learning techniques and use the encoder part as the perceptual loss neural network (L_{AE}), which has been introduced in section II-B. The discriminator D is a 6-layer convolutional neural network with 64, 64, 128, 128, 256, and 256 filters in each layer, followed by two fully-connected layers, of which the first layer has 1024 outputs and the last layer has a single output.

III. Experiments

In this section, we introduce the dataset used to train and evaluate the networks and discussing the experimental setup, including data preparation and selection of hyperparameters. We present the network's performance in denoising CT images and compare our method to recent state-of-the-art methods. Our experiments show that the proposed self-attention CNN and autoencoder perceptual loss is beneficial to neural networks both with or without WGAN. The following experimental results are divided into with-WGAN and without-WGAN versions.

A. Data Set

We use a publicly available dataset released as part of the *2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge*. [1] This dataset contains 10 anonymized patient normal-dose abdominal CT images (acquired at 120 kV and 200 effective mAs), and Poisson noise has been inserted into the projection data for each case to reach a noise level that corresponds to quarter-dose levels. The images in each case are reconstructed with a medium smooth kernel (B30 kernel) and a medium sharp kernel (D45 kernel), respectively. For each case, slice thicknesses of 3mm and 1mm were provided. In this work, we utilize the scans reconstructed using the B30 kernel and 1 mm slice thickness. For training, we use the quarter-dose CT images as input and use the NDCT images as the training target. A total of *25600* pairs of CT image patches are randomly selected from 6 patients' scan in the dataset for training, *1382 and 1345* pairs of image slices from 4 patients' scan are set aside for validation and test, respectively. Each patch is sized of 64×64. We stack 3 patches to form a 3D input to our neural network. Normalization of the input patches is performed to scale the pixel values of CT images to values between zero and one.

B. Network Training

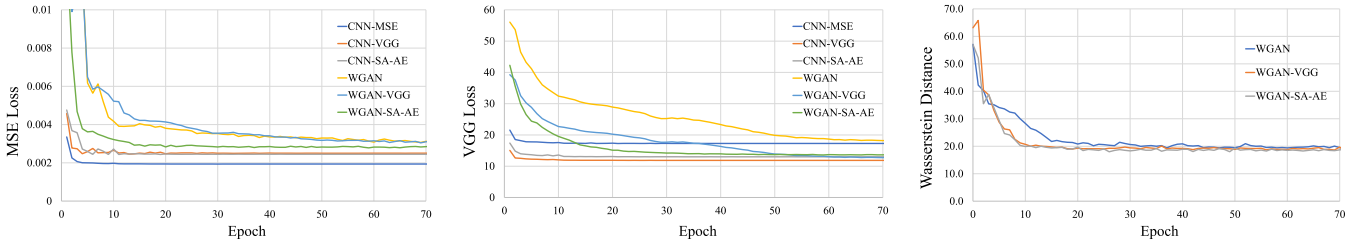
During training, we use Adam [59] to optimize all of the networks. The hyper parameters for Adam are set as follows: the learning rate is $\alpha = 1.0 \times e^{-4}$ and the two exponential decay rates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For WGAN training, the weighting parameter λ that controls the tradeoff between Wasserstein distance and gradient penalty is set to 10. The weighting parameter η that controls the tradeoff between WGAN loss and perceptual loss is set as 0.1 as suggested in [53]. The networks are implemented using Pytorch, and are trained/validated on a workstation with a NVIDIA Tesla P100 GPU.

C. Convergence

We perform experiments on different neural networks with various settings for comparison, which are listed in Table I. In order to show the effectiveness of self-attention and autoencoder, we train neural networks both *with* and *without* WGAN. For the networks *without* WGAN, we compare the proposed network CNN-SA-AE with CNN-MSE and CNN-VGG networks. For the networks *with* WGAN, we compare the proposed network WGAN-SA-AE with WGAN and WGAN-VGG networks.

Fig. 6 visualizes the training procedure and convergence of the aforementioned neural networks, which are measured with

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



(a) MSE loss (b) VGG loss (c) Wasserstein distance

Fig. 6. Network convergence over different loss functions. (a) MSE loss convergence, (b) VGG loss convergence, and (c) Wasserstein estimation convergence.

TABLE I
LIST OF ALL TRAINED NETWORKS

Experiment name	Generator network	Loss function	with WGAN
CNN-MSE	CNN	L_{MSE}	no
CNN-VGG	CNN	L_{VGG}	no
CNN-AE	CNN	L_{AE}	no
SACNN-MSE	Self-attention CNN	L_{MSE}	no
SACNN-VGG	Self-attention CNN	L_{VGG}	no
SACNN-AE	Self-attention CNN	L_{AE}	no
WGAN	CNN	L_{WGAN}	yes
WGAN-VGG	CNN	$L_{WGAN} + L_{VGG}$	yes
WGAN-SA-AE	Self-attention CNN	$L_{WGAN} + L_{AE}$	yes

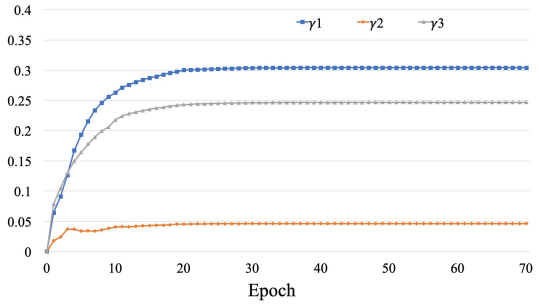
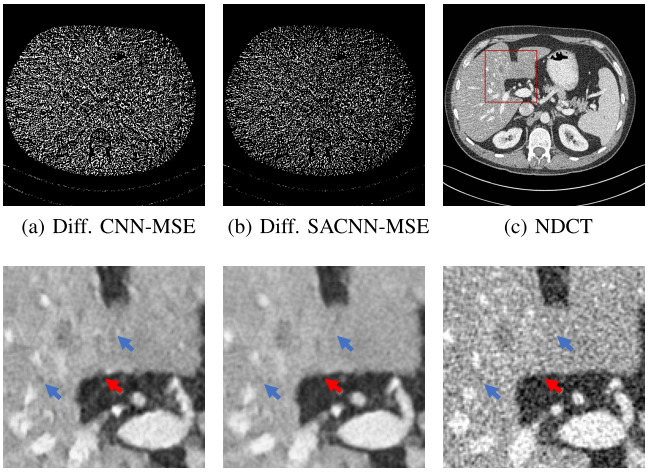


Fig. 7. Training curves of the scaling factor γ in attention layers.



(a) Diff. CNN-MSE (b) Diff. SACNN-MSE (c) NDCT

(d) CNN-MSE (e) SACNN-MSE (f) NDCT

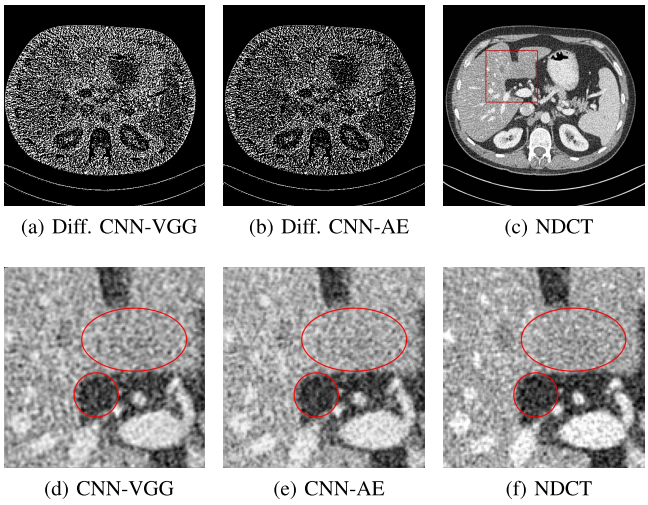
Fig. 8. Ablation study of self-attention (SA) by comparing denoised CT images generated by CNNs *without* self-attention (CNN-MSE) and *with* self-attention (SACNN-MSE). (a) difference image between the NDCT image and the denoised image of CNN-MSE. (b) difference image between the NDCT image and the denoised image of SACNN-MSE. (c) and (f) ground-truth normal dose CT image and zoomed ROI of the red rectangle. (d) zoomed ROI of CNN-MSE's denoised image. (e) zoomed ROI of SACNN-MSE's denoised image.

MSE loss, VGG loss, and Wasserstein distance. Fig. 6a shows the average MSE loss versus the number of epochs over the validation set. From this figure, we observe that the proposed WGAN-SA-AE network not only converges faster than other WGAN-based counterparts but also achieves lower MSE loss. Due to its simplicity, all three CNN designs *without* training on WGAN converge rapidly. Since CNN-MSE is using MSE-loss as its objective, it achieves the lowest error in terms of MSE-loss. However, it suffers from oversmoothing. Fig. 6b shows the convergence curve of VGG-loss, in which the proposed WGAN-SA-AE has the fastest convergence among all WGAN-based methods here. The VGG loss of WGAN-SA-AE is very close to WGAN-VGG, and the latter one achieves lower VGG-loss because VGG-loss is WGAN-VGG's native training objective. CNN-VGG achieves the lowest VGG-loss. However, the lower VGG loss does not guarantee better performance. Our experiment shows that noises generated by VGG-loss network weakens its advantage; this observation will be further discussed in the following sections. We show convergence curve of Wasserstein distance in Fig. 6c, where the WGAN-SA-AE takes the shortest time to converge and achieves slightly better score in terms of the Wasserstein distance than the state-of-the-art WGAN-based counterparts.

As described in (5), we apply a trainable parameter γ on attention outputs. Fig. 7 shows the training curve of γ^1, γ^2, γ^3 in three attention blocks respectively. They are initialized as zeros, and grow larger in the training procedure, and finally flatten out at around 20^{th} epoch.

D. Ablation Study of Proposed Methods

In this section, we perform model ablation studies to validate the effectiveness of the proposed self-attention (SA) module and autoencoder (AE) module independently. We show the subjective results from Fig. 8 to 9. We also show the objective results and summarize in Table II. In this table, we evaluate the mean and variance of peak-to-noise ratio (PSNR) and structural similarity index (SSIM) values on the test set, comprised of a total of 1345 slices across two CT studies.In our experiments, the normal dose CT images are used as the reference.



(a) Diff. CNN-VGG (b) Diff. CNN-AE (c) NDCT

(d) CNN-VGG (e) CNN-AE (f) NDCT

Fig. 9. Ablation study of autoencoder (AE) loss by comparing denoised CT images generated by CNNs *with* VGG loss (CNN-VGG) and *with* AE loss (CNN-AE). (a) difference image between the NDCT image and the denoised image of CNN-VGG. (b) difference image between the NDCT image and the denoised image of CNN-AE. (c) and (f) ground-truth normal dose CT image and zoomed ROI of the red rectangle. (d) zoomed ROI of CNN-VGG's denoised image. (e) zoomed ROI of SACNN-AE's denoised image.

TABLE II
THE OBJECTIVE TEST RESULT OF DIFFERENT NETWORKS LIST IN TABLE I

Method	PSNR		SSIM	
	MEAN	STD	MEAN	STD
LDCT	22.0123	2.4288	0.7845	0.0652
CNN-MSE	27.6842	2.1564	0.8841	0.0489
CNN-VGG	24.2674	2.8951	0.8040	0.0617
CNN-AE	24.9852	2.2403	0.8373	0.0623
SACNN-MSE	**27.7371**	2.5374	**0.8876**	0.0584
SACNN-VGG	26.3797	2.9378	0.8671	0.0680
SACNN-AE	26.6020	2.0415	0.8784	0.0501
WGAN	23.4134	1.9476	0.8047	0.0739
WGAN-VGG	22.9558	1.7955	0.7909	0.0504
WGAN-SA-AE	23.9394	2.3787	0.8403	0.0569

To validate the effectiveness of the self-attention module, we compare the denoised images that generated by CNN *without* self-attention and by CNN *with* self-attention, as is shown in Fig. 8. Both networks are trained with MSE loss. We annotate them as CNN-MSE and SACNN-MSE respectively. Fig. 8c and Fig. 8f show the normal dose CT image and the zoomed ROI as ground truth. Compared to pure CNN-MSE without self-attention, SACNN avoids some waxy artifacts and looks smoother at the surface of organs, as is pointed by the *blue* arrows in Fig. 8d and Fig. 8e. SACNN also generates sharper edges as indicated by the *red* arrows. To compare the denoised images generated by CNN-MSE and SACNN-MSE, we visualize the difference images between NDCT image and each denoised image, as is shown in Fig. 8a and Fig. 8b. We see from the difference images that our self-attention module avoids more noise, which indicates that SACNN-MSE is closer to the NDCT image. Table II also shows the advantage of self-attention, where both the PSNR and SSIM of SACNN-MSE are better than those of the CNN-MSE.

TABLE III
OBJECTIVE PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS. THESE FOUR METRICS ARE PEAK-TO-NOISE RATIO (PSNR), STRUCTURAL SIMILARITY (SSIM), INFORMATION FIDELITY CRITERION (IFC), AND VISUAL INFORMATION FIDELITY (VIF)

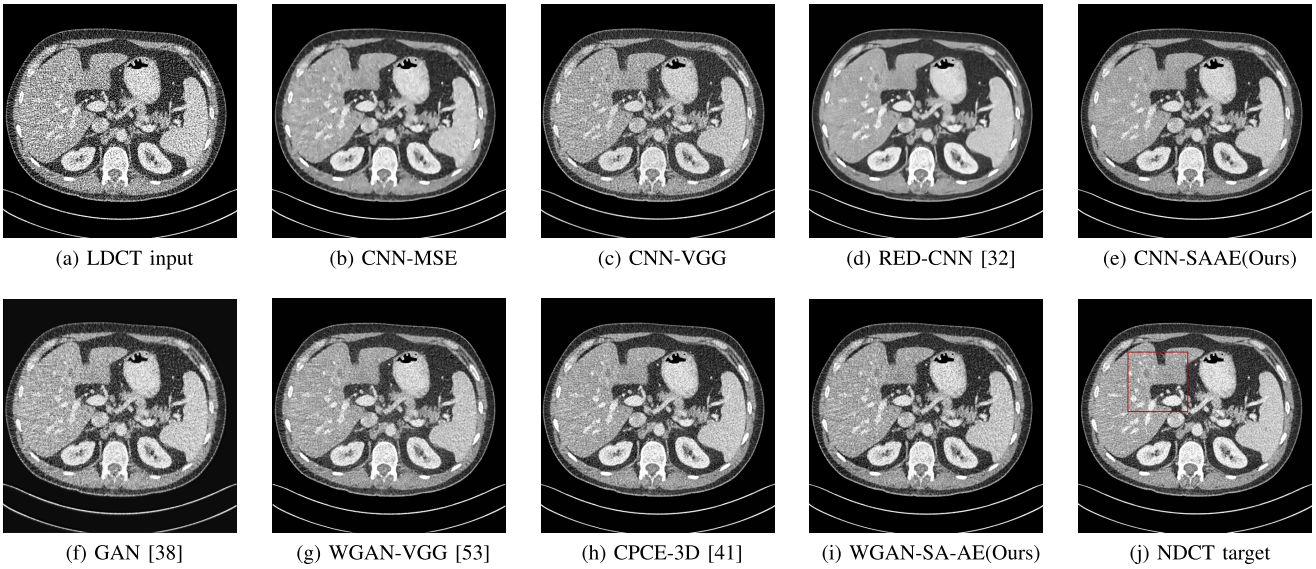
Method	Test set			
	PSNR	SSIM	VIF	IFC
LDCT	19.5978	0.7270	0.1758	1.1730
CNN-MSE	22.2436	0.7472	0.1844	1.1945
CNN-VGG	21.8586	0.7165	0.1970	1.2976
RED-CNN	**22.6464**	**0.7839**	0.2001	1.4208
CNN-SA-AE (Ours)	22.2789	0.7816	**0.2071**	**1.4258**
GAN	20.9463	0.7366	0.1355	0.8717
CPCE-3D	22.1429	0.7795	0.2015	1.3588
WGAN-VGG	21.9328	0.7708	0.1897	1.2726
WGAN-SA-AE (Ours)	**22.1758**	**0.7800**	**0.2057**	**1.3899**

To validate the effectiveness of the auto-encoder module, we compare the CNNs trained *with* VGG perceptual loss and *with* autoencoder perceptual loss, as is shown in Fig. 9. We annotate them as CNN-VGG and CNN-AE respectively. By comparing Fig. 9e to Fig. 9d, we can see that the textures generated by AE-loss network is finer than that of VGG-loss network, as is shown in the circled areas. The difference images in Fig. 9a and Fig. 9b show that the denoised image generated by CNN-AE network is less cloudy, which indicates that Fig. 9e is more similar to the ground-truth image. In Table II, CNN-AE also achieves higher PSNR and SSIM than CNN-VGG network.

In addition, by comparing Fig. 9a and Fig. 9b to Fig. 8a and Fig. 8b, we can also find that the former two show images more structural shapes than the latter two. We posit that this difference exists because the latter two are trained with MSE loss, which tends to minimize pixel-level similarity. However, the perceptual loss, such as VGG and AE, are more focused on optimizing feature-level similarity. In the next sub-section, we apply both the self-attention and autoencoder on CNN-based and WGAN-based network, and show the comparison results with CNN-MSE, CNN-VGG and some state-of-the-art methods.

E. Denoising Result Comparison

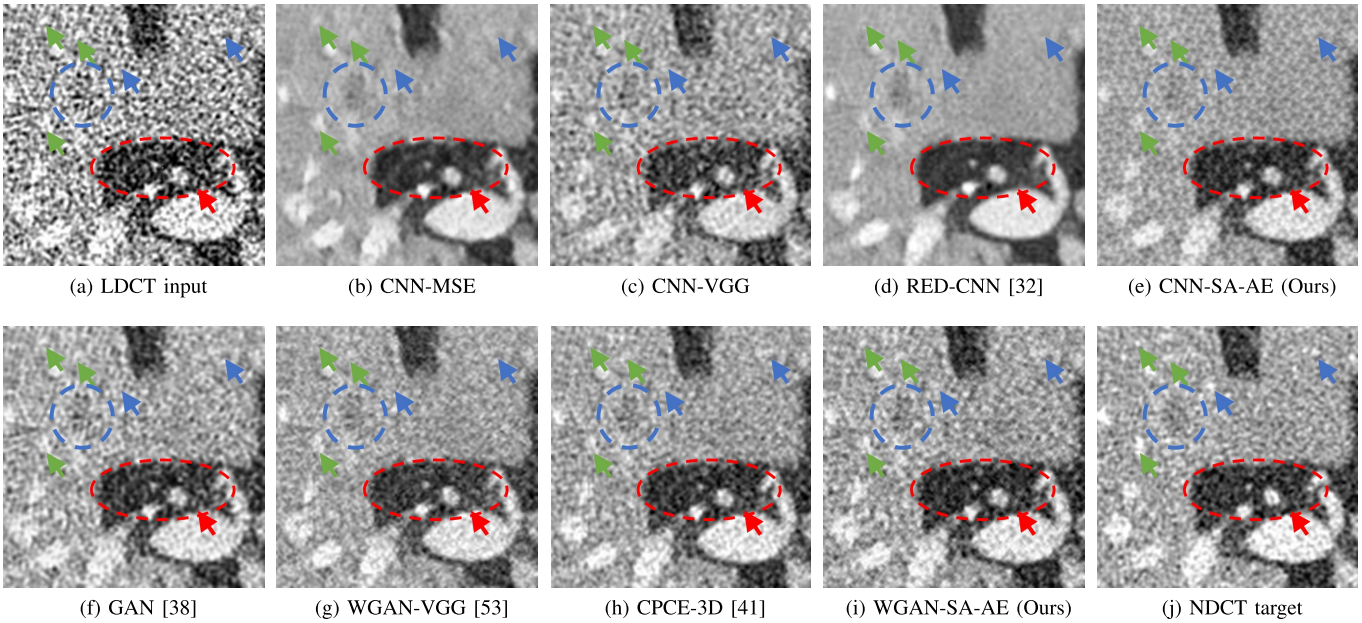
To visualize the denoising performance, we depict two patients' scans from the test set, containing anatomy with a lesion. We depict the results of CNN-MSE, CNN-VGG, our proposed CNN-SA-AE and WGAN-SA-AE networks. For comparison, we also show the results of several state-of-the-art methods, including RED-CNN [32], GAN [38], WGAN-VGG [53], and CPCE-3D [41]. Fig. 10 and Fig. 12



(a) LDCT input (b) CNN-MSE (c) CNN-VGG (d) RED-CNN [32] (e) CNN-SAAE(Ours)

(f) GAN [38] (g) WGAN-VGG [53] (h) CPCE-3D [41] (i) WGAN-SA-AE(Ours) (j) NDCT target

Fig. 10. Performance comparison of state-of-the-art methods on case L291 from the NIH-AAPM-Mayo Clinic dataset. The display window is [−180, 200] HU.



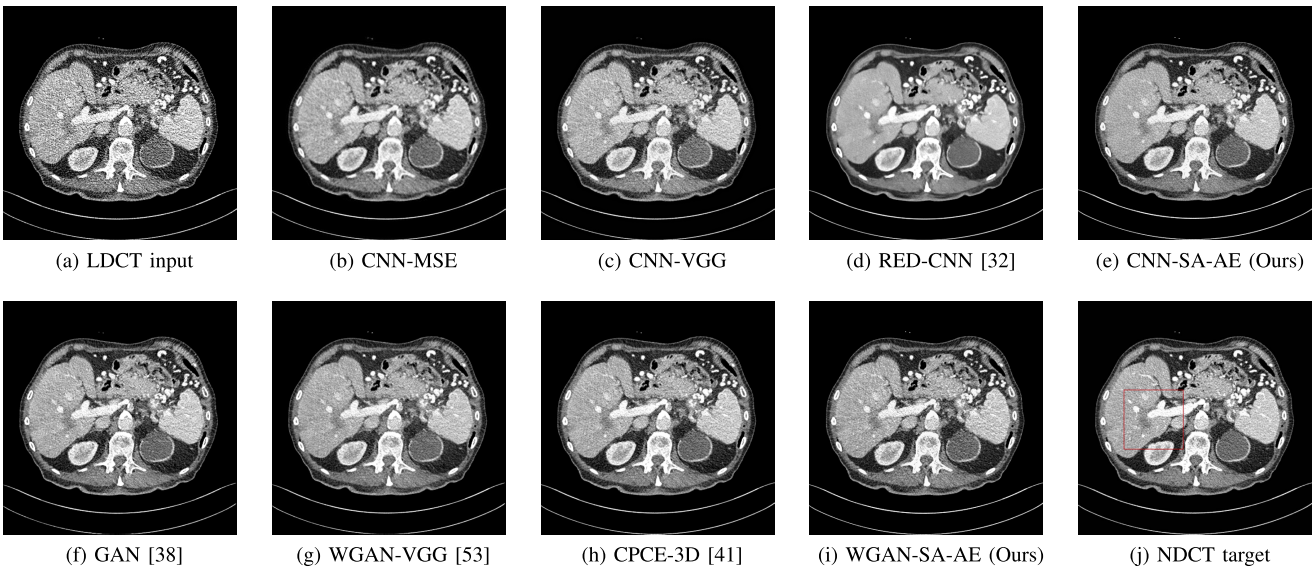
(a) LDCT input (b) CNN-MSE (c) CNN-VGG (d) RED-CNN [32] (e) CNN-SA-AE (Ours)

(f) GAN [38] (g) WGAN-VGG [53] (h) CPCE-3D [41] (i) WGAN-SA-AE (Ours) (j) NDCT target

Fig. 11. Zoomed ROI of the red rectangle in Fig. 10. The blue circle indicates low attenuation lesion in the left lobe of liver, the red ellipse indicates the background region, and the arrows indicates vessels. The display window is [−180, 200] HU.

show the visualization results on two representative slices in case L291 and case L506 respectively, and Fig. 11 and Fig. 13 show the enlarged regions of interest (ROIs) marked by the red rectangles in Fig. 10 and Fig. 12. For fair comparison, we align our display window settings with CPCE [41], where Fig. 10 is [−180, 200] HU and Fig. 12 is [−160, 240] HU.

Fig. 10j is the normal dose CT (NDCT), and Fig. 10a is quarter-dose CT (LDCT). Fig. 10b to Fig. 10e show the denoising results of CNNs without GAN framework, including CNN-MSE, CNN-VGG, RED-CNN and CNN-SA-AE. Fig. 10f to Fig. 10i show the denoising results of GAN-based frameworks, including GAN, CPCE-3D, WGAN-VGG and WGAN-SA-AE. In above figures, all networks show their capability of generating denoised images. In the following paragraphs, we illustrate their differences with the pointing arrows in detail. Fig. 10b and Fig. 10d show the denoising result of CNN-MSE and RED-CNN, which are trained with MSE loss. From their enlarged ROIs shown in Fig. 11b and Fig. 11d, we see that they have good denoising performance but lose some fine structural objects (vessels) as is pointed out by the *green* arrows, compared to NDCT in Fig. 11j. This is because MSE loss are focused on minimizing the pixel-level average loss, which often generates over-smooth results. Previous work [41], [53] use VGG perceptual loss to avoid this problem. Similar improvements are also observed in our experiments by comparing Fig. 11c to Fig. 11b.



(a) LDCT input (b) CNN-MSE (c) CNN-VGG (d) RED-CNN [32] (e) CNN-SA-AE (Ours)

(f) GAN [38] (g) WGAN-VGG [53] (h) CPCE-3D [41] (i) WGAN-SA-AE (Ours) (j) NDCT target

Fig. 12. Performance comparison of state-of-the-art methods on case L506 from the NIH-AAPM-Mayo Clinic dataset. The display window is [−160, 240] HU.

TABLE IV
SUBJECTIVE QUALITY SCORE (MEAN ± SD) FOR DIFFERENT METHODS

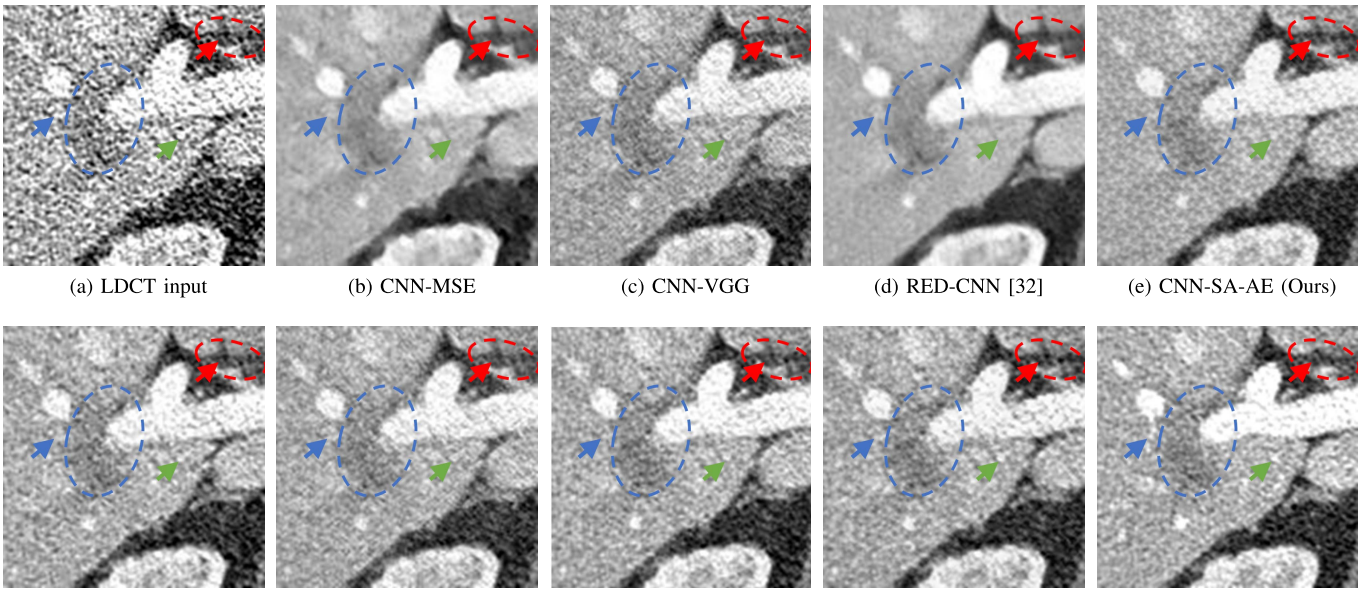
	NDCT	LDCT	CNN-MSE	CNN-VGG	RED-CNN	CNN-SA-AE	WGAN-VGG	CPCE-3D	WGAN-SA-AE
Noise Reduction	-	-	3.21±0.13	2.81±0.19	**3.58±0.20**	3.35±0.13	3.12±0.19	3.20±0.15	3.33±0.19
Artifact Reduction	-	-	2.87±0.31	3.37±0.20	3.12±0.28	3.45±0.16	3.57±0.11	3.65±0.19	**3.67±0.20**
Structure Preservation	-	-	3.01±0.22	3.31±0.16	3.18±0.19	3.40±0.21	3.35±0.18	3.45±0.27	**3.46±0.17**
Overall Quality	4.00±0.09	1.05±0.09	3.07±0.17	3.21±0.14	3.29±0.23	3.41±0.18	3.43±0.10	3.51±0.14	**3.58±0.21**

But the image generated by VGG-loss in Fig. 11c contains more noises than RED-CNN, which may weaken its strength. The proposed self-attention and auto-encoder modules help to solve this problem by realizing a better balance between noise reduction and structure preservation, as is shown in Fig. 10e. Similar improvements can also be witnessed in WGAN-based networks, as is pointed out by the colored arrows and circles in Fig. 11i. At the positions pointed by the three green arrows, vessels in Fig. 11g and Fig. 11f look more or less vague and sometimes vanished whereas they are clearly identifiable in Fig. 11i. The Fig. 11i generated by WGAN-SA-AE is most similar and comparable to Fig. 11h generated by the state-of-the-art CPCE. The vessels pointed by the three green arrows in WGAN-SA-AE are slightly brighter than those in CPCE. In addition, the WGAN-SA-AE approach has the advantage of noise reduction over previous methods, especially at dark background area in the *red* circle and pointed by the *red* arrows from Fig. 11f to Fig. 11i. The *blue* arrows point out some shape deformation in all the denoised images. They are almost smeared out in most networks, especially in CNN-MSE, RED-CNN, and WGAN-VGG. But CPCE and WGAN-SA-AE have minor responses. We think this is acceptable because the signals at corresponding positions in the LDCT input are weak compared to the noises in the neighborhood. We do not want to generate strong outputs from weak inputs. The low attenuation lesion, indicated by the *blue* circle, is clearly visible in all methods here. Above observations can also be witnessed in the corresponding denoised results of sample 506, shown in Fig. 12 and Fig. 13.

F. Quantitative Analysis

We calculate four full-reference quality assessments, peak-to-noise ratio (PSNR), structural similarity (SSIM), information fidelity criterion (IFC) [60], and visual information fidelity (VIF) [61]. Table III summarizes the comparison results tested on the test set. We compare the results with other state-of-the-art methods, including RED-CNN [32], GAN [38], WGAN-VGG [53] and CPCE-3D [41].

RED-CNN achieves the highest PSNR as a result of the MSE loss function. However, the MSE-based approach has the problem of over-smooth and lose fine structural information as is shown in Fig. 10 and Fig. 12. CNN-VGG has lower PSNR and SSIM values but it preserves more structural details as is discussed in previous section. The proposed CNN-SA-AE achieves the best VIF and IFC, and the second best PSNR and SSIM among the four methods using CNN-based approaches. The denoised images in Fig. 11e and Fig. 13e also show that the network enhanced by self-attention and auto-encoder achieves a better balance of denoising and structure preservation. Among the four metrics, PSNR and SSIM have more focus on pixel-level similarity, VIF and IFC have more focus on psychovisual features of the human visual system by using natural statistics models.



(a) LDCT input (b) CNN-MSE (c) CNN-VGG (d) RED-CNN [32] (e) CNN-SA-AE (Ours)

(f) GAN [38] (g) WGAN-VGG [53] (h) CPCE-3D [41] (i) WGAN-SA-AE (Ours) (j) NDCT target

Fig. 13. Zoomed ROI of the red rectangle in Fig. 12. The blue ellipse indicates low attenuation lesion in the posterior right liver lobe, the red ellipse indicates the background region, and the arrows indicates vessels. The display window is [−160, 240] HU.

Meanwhile, for the GAN-based approaches, GAN network achieves the worst objective values on these four metrics. Both CPCE-3D and WGAN-VGG use Wasserstein Distance and VGG perceptual loss to improve the GAN network, and achieve better results on the objective metrics. The proposed WGAN-SA-AE inherits the advantage of the WGAN framework, and receives more information with long-range dependencies by using self-attention block. Our proposed WGAN-SA-AE have slightly better scores than state-of-the-art works. We can see from the result that the proposed self-attention and autoencoder perceptual loss can improve the denoising performance of both GAN-based and pure CNN-based networks.

G. Blind Reader Study

We randomly select 10 groups of image slices in the test set and perform a blind reader study. Each group contains LDCT image, NDCT image, and denoised iamges, including CNN-MSE, CNN-VGG, RED-CNN [32], CPCE-3D [41], WGAN-VGG [53], CNN-SA-AE, and WGAN-SA-AE. In each group, different methods are blind to radiologists, except for LDCT and NDCT which are labeled as the reference images. Two radiologists were asked to score each image slice independently in terms of noise reduction, artifact reduction, structure preservation, and overall quality on the five-point scale (1 = unacceptable and 5 = excellent). For NDCT and LDCT images, only the overall quality was scored as the golden standard. For each method, the mean and standard deviation values are calculated from two radiologists times 10 images as the final results. As is shown in Table IV, MSE-based approaches achieve the best noise reduction performance, GAN-based and networks with perceptual loss achieve better score in terms of structure preservation and artifact reduction. The proposed CNN-SA-AE network has better balance between noise reduction and structure preservation. Both the WGAN-VGG, CPCE-3D and the proposed WGAN-SA-AE are good at structure preservation, but WGAN-SA-AE generates less noises, which is also seen in Fig. 10 and Fig. 12. In summary, MSE-based network is good at noise reduction while losing some structure details, which result in the image degradation for diagnosis. Perceptual loss and GAN framework are good at preserving structure details. With the help of self-attention and antoencoder module, the proposed WGAN-SA-AE offers a better balance between noise reduction and structure preservation, and finally gets better overall quality.

IV. Discussion and Conclusion

This paper introduces a novel low-dose CT denoising approach. The motivation for this paper includes two parts: 1) The traditional convolution operation equally convolves a certain region of pixels on input CT images, and thus it is inefficient to capture the global information across the whole CT image. 2) The typically perceptual loss adopts the VGG network which is trained on a natural image dataset, which causes concerns on how well it performs on CT image feature extraction. The contributions of this work are as follows: 1) We introduce a 3D self-attention mechanism on convolution block, which brings regional attention for feature filtering and augments the capability of convolution for capturing global information. 2) we propose to use a self-supervised learning scheme to train an auto encoder network. Trained and tested on the dataset of *2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge*, we show that the proposed methods can efficiently enhance traditional CNNs and can achieve comparable or sometimes better results than the state-of-the-art methods. The proposed self-attention and self-supervised auto encoder perceptual loss are demonstrated to be beneficial

for both approaches with meaningful improvement on training convergence and denoising results.

In the quantitative results comparison, our method is not the best in some cases when using PSNR and SSIM. This is because we do not use pixel-wise loss function and pay more attention on structural features. Despite the improvements from our methods, the generated images still contain some noises as well as some fine structure deformation compared to the original NDCT image. For future work, we will consider to investigate the design space of different CNN architectures with self-attention and auto-encoder modules. We will also consider to evaluate the model using locally collected raw CT data, and to evaluate the impact on radiomic features. Given that measures such as PSNR and SSIM are not ideal to measures improvement in image quality, we will investigate how the proposed normalization approach affects the performance of target tasks such as nodule detection and characterization.

REFERENCES

[1] R. Smith-Bindman *et al.*, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer," *Arch. Int. Med.*, vol. 169, no. 22, pp. 2078–2086, 2009.

[2] A. B. de González *et al.*, "Projected cancer risks from computed tomographic scans performed in the united states in 2007," *Arch. Int. Med.*, vol. 169, no. 22, pp. 2071–2077, 2009.

[3] 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, Oct. 2009.

[4] M. Balda, J. Hornegger, and B. Heismann, "Ray contribution masks for structure adaptive Sinogram filtering," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1228–1239, Jun. 2012.

[5] 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.

[6] T. Li *et al.*, "Nonlinear Sinogram smoothing for low-dose X-ray CT," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 5, pp. 2505–2513, Oct. 2004.

[7] 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, pp. 4777–4807, Sep. 2008.

[8] 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, pp. 5949–5967, Sep. 2011.

[9] 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, pp. 7923–7956, Dec. 2012.

[10] Y. Zhang, W. Zhang, Y. Lei, and J. Zhou, "Few-view image reconstruction with fractional-order total variation," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 31, no. 5, p. 981, May 2014.

[11] Y. Chen *et al.*, "Bayesian statistical reconstruction for low-dose X-ray computed tomography using an adaptive-weighting nonlocal prior," *Comput. Med. Imag. Graph.*, vol. 33, no. 7, pp. 495–500, Oct. 2009.

[12] J. Ma *et al.*, "Iterative image reconstruction for cerebral perfusion CT using a pre-contrast scan induced edge-preserving prior," *Phys. Med. Biol.*, vol. 57, no. 22, pp. 7519–7542, Nov. 2012.

[13] Y. Zhang, Y. Xi, Q. Yang, W. Cong, J. Zhou, and G. Wang, "Spectral CT reconstruction with image sparsity and spectral mean," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 510–523, Dec. 2016.

[14] 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, Sep. 2012.

[15] 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.

[16] M. Yan, J. Chen, L. A. Vese, J. Villasenor, A. Bui, and J. Cong, "EM+TV based reconstruction for cone-beam CT with reduced radiation," in *Proc. Int. Symp. Vis. Comput.* Berlin, Germany: Springer, 2011, pp. 1–10.

[17] J. Adler and O. Oktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, Jun. 2018.

[18] K. Hammernik, T. Würfl, T. Pock, and A. Maier, "A deep learning architecture for limited-angle computed tomography reconstruction," in *Bildverarbeitung für die Medizin*. Berlin, Germany: Springer, 2017, pp. 92–97.

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

[20] Z. Li *et al.*, "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Med. Phys.*, vol. 41, no. 1, Dec. 2013, Art. no. 011908.

[21] Z. S. Kelm, D. Blezek, B. Bartholmai, and B. J. Erickson, "Optimizing non-local means for denoising low dose CT," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro*, Jun. 2009, pp. 662–665.

[22] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[23] 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, pp. 5803–5820, Aug. 2013.

[24] 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, pp. 5401–5415, Sep. 2010.

[25] K. Sheng, S. Gou, J. Wu, and S. X. Qi, "Denoised and texture enhanced MVCT to improve soft tissue conspicuity," *Med. Phys.*, vol. 41, no. 10, Oct. 2014, Art. no. 101916.

[26] D. Kang *et al.*, "Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm," *Proc. SPIE, Med. Imag., Image Process.*, vol. 8669, Mar. 2013, Art. no. 86692G.

[27] A. Mendrik, E.-J. Vonken, A. Rutten, M. Viergever, and B. Van Ginneken, "Noise reduction in computed tomography scans using 3-D anisotropic hybrid diffusion with continuous switch," *IEEE Trans. Med. Imag.*, vol. 28, no. 10, pp. 1585–1594, Oct. 2009.

[28] 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.

[29] M. Li, S. Shen, W. Gao, W. Hsu, and J. Cong, "Computed tomography image enhancement using 3D convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 291–299.

[30] H. Chen *et al.*, "Low-dose CT via convolutional neural network," *Biomed. Opt. Express*, vol. 8, no. 2, pp. 679–694, 2017.

[31] D. Wu, K. Kim, G. E. Fakhri, and Q. Li, "A cascaded convolutional neural network for X-ray low-dose CT image denoising," 2017, *arXiv:1705.04267*. [Online]. Available: https://arxiv.org/abs/1705.04267

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

[33] 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, pp. e360–e375, Oct. 2017.

[34] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[36] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

[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] C. You *et al.*, "Structurally-sensitive multi-scale deep neural network for low-dose CT denoising," *IEEE Access*, vol. 6, pp. 41839–41855, 2018.

[40] X. Yi and P. Babyn, "Sharpness-aware low-dose CT denoising using conditional generative adversarial network," *J. Digit. Imag.*, vol. 31, no. 5, pp. 655–669, Oct. 2018.

[41] H. Shan *et al.*, "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1522–1534, Jun. 2018.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[43] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318
[44] T. Miyato and M. Koyama, "cGANs with projection discriminator," 2018, *arXiv:1802.05637*. [Online]. Available: https://arxiv.org/abs/1802.05637
[45] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2005, pp. 60–65.
[46] 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, Aug. 2007.
[47] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1033–1038.
[48] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 349–356.
[49] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 24.
[50] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[51] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 373–389.
[52] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[53] Q. Yang *et al.*, "Low-dose CT image Denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.
[54] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising autoencoders: Learning useful representations in a deep network with a local Denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1953039
[55] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, Sep. 2014.
[56] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[57] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," 2016, *arXiv:1606.01933*. https://arxiv.org/abs/1606.01933
[58] 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.*, Jun. 2009, pp. 248–255.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[60] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, 2005.
[61] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Jan. 2004, p. iii-709.