Structure-preserving low-dose computed tomography image denoising using a deep residual adaptive global context attention network

Yuanke Zhang[1], Dejing Hao[1], Yingying Lin[1], Wanxin Sun[1], Jinke Zhang[1], Jing Meng[1], Fei Ma[1], Yanfei Guo[1], Hongbing Lu[2], Guangshun Li[1], Jianlei Liu[1]

[1]School of Computer Science, Qufu Normal University, Rizhao, China; [2]School of Biomedical Engineering, Fourth Military Medical University, Xi'an, China

Contributions: (I) Conception and design: Y Zhang; (II) Administrative support: Y Zhang, J Liu; (III) Provision of study materials or patients: J Meng, H Lu, G Li; (IV) Collection and assembly of data: Y Lin, W Sun, J Zhang; (V) Data analysis and interpretation: D Hao, F Ma, Y Guo, J Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Yuanke Zhang, PhD; Jianlei Liu, PhD. School of Computer Science, Qufu Normal University, No. 80, Yantai Road, Rizhao 276826, China. Email: yuankezhang@163.com; jianleiliu@qfnu.edu.cn.

Background: Low-dose computed tomography (LDCT) scans can effectively reduce the radiation damage to patients, but this is highly detrimental to CT image quality. Deep convolutional neural networks (CNNs) have shown their potential in improving LDCT image quality. However, the conventional CNN-based approaches rely fundamentally on the convolution operations, which are ineffective for modeling the correlations among nonlocal similar structures and the regionally distinct statistical properties in CT images. This modeling deficiency hampers the denoising performance for CT images derived in this manner.

Methods: In this paper, we propose an adaptive global context (AGC) modeling scheme to describe the nonlocal correlations and the regionally distinct statistics in CT images with negligible computation load. We further propose an AGC-based long-short residual encoder-decoder (AGC-LSRED) network for efficient LDCT image noise artifact-suppression tasks. Specifically, stacks of residual AGC attention blocks (RAGCBs) with long and short skip connections are constructed in the AGC-LSRED network, which allows valuable structural and positional information to be bypassed through these identity-based skip connections and thus eases the training of the deep denoising network. For training the AGC-LSRED network, we propose a compound loss that combines the L_1 loss, adversarial loss, and self-supervised multi-scale perceptual loss.

Results: Quantitative and qualitative experimental studies were performed to verify and validate the effectiveness of the proposed method. The simulation experiments demonstrated the proposed method exhibits the best result in terms of noise suppression [root-mean-square error (RMSE) =9.02; peak signal-to-noise ratio (PSNR) =33.17] and fine structure preservation [structural similarity index (SSIM) =0.925] compared with other competitive CNN-based methods. The experiments on real data illustrated that the proposed method has advantages over other methods in terms of radiologists' subjective assessment scores (averaged scores =4.34).

Conclusions: With the use of the AGC modeling scheme to characterize the structural information in CT images and of residual AGC-attention blocks with long and short skip connections to ease the network training, the proposed AGC-LSRED method achieves satisfactory results in preserving fine anatomical structures and suppressing noise in LDCT images.

Keywords: Low-dose computed tomography (LDCT); image denoising; deep residual network; adaptive global context attention (AGC attention); multiscale perceptual loss

Introduction

With a high capability to image the internal structure of the human body in a noninvasive manner, computed tomography (CT) is critical in detecting lesions, tumors, and metastasis (1). However, the high level of accumulated radiation exposure from a CT examination and the risk of radiation-induced cancer and genetic or other diseases is of a significant concern to patients and operators. Minimizing X-ray exposure to patients has been one of the major efforts undertaken in the CT field (2,3). As the tube current [milliampere-seconds (mAs)] is linearly related to the radiation dose, a reduction in mAs is perhaps the simplest and most effective way to reduce radiation exposure. However, the low-mAs-acquisition protocols may be highly detrimental to image quality, resulting in images with unavoidable noise-induced artifacts, which may hamper detection accuracy.

Thus far, various noise suppression strategies have been proposed to address the noise artifacts problem in low-dose CT (LDCT), including sinogram domain smoothing (4,5), model-based iterative reconstruction (MBIR) (6,7), and image domain denoising (8,9). Sinogram domain smoothing methods seek an optimal estimation of the ideal projection by optimizing a cost function in sinogram domain and then reconstructs the CT image from the estimated projection via the traditional filtered back-projection (FBP) algorithm. MBIR methods optimize a cost function according to both the raw data statistics and the prior knowledge of the reconstructed object for image reconstruction by using iterative algorithms. Although these above methods can suppress the noise of CT images, they depend heavily on the manual design of appropriate prior models, posing a significant challenge to researchers. In addition, the CT images reconstructed with these techniques still suffer from the oversmoothing of subtle tissue structures. Image domain denoising methods are postprocessing techniques which mitigate noise and artifacts directly from reconstructed CT images. Conventional postprocessing methods include nonlocal mean algorithms (10,11), dictionary-learning-based algorithms (9,12), low-rank algorithms (13,14), and diffusion filter algorithms (15), among others. Since the noise artifact statistics in the reconstructed LDCT images are inhomogeneous, using these conventional postprocessing methods to achieve a good balance between fine structure preservation and noise artifact suppression is difficult.

Recently, with the rapid development of deep learning techniques, deep convolutional neural networks (CNNs), which learn nonlinear parametric mapping from a low-quality data manifold to a high-quality data manifold, have shown considerable potential for LDCT image noise suppression. For example, Chen *et al.* (16) combined an autoencoder, deconvolution network, and shortcut connections with a residual encoder-decoder CNN (RED-CNN) for LDCT imaging. Yang *et al.* (17) proposed a new CT image denoising method based on the generative adversarial network with Wasserstein distance and perceptual similarity. Zavala-Mondragon *et al.* (18) proposed a learned wavelet-frame shrinkage network (LWFSN) and its residual counterpart (rLWFSN) for LDCT image noise suppression.

Tissue structures in CT images show evident nonlocal self-similarity properties (19,20). The global contextual information across large tissue regions, otherwise known as long-range dependency, is desirable for modeling the correlations among nonlocal similar structures. On the other hand, the conventional CNN-based approaches are based fundamentally on the convolution operations. They extract informative features within local receptive fields; thus, the global contextual information can only be captured by deeply stacking a series of convolutional layers. However, a deeper network architecture suffers from optimization difficulty and computational inefficiency. The pooling layers may increase the size of the receptive fields of the CNN networks, but the simple maximizing or averaging feature aggregation strategy hinders its representational ability for meaningful global contextual information. The nonlocal network (NLnet) (21), however, solves this problem via a self-attention mechanism. For each query position, the NLnet computes the query-specific global context (GC) as a weighted sum of the features at all positions in the input feature images to guide the convolutional filtering. For example, Li *et al.* (22) proposed a novel three-dimensional (3D) self-attention CNN for the LDCT denoising problem. Bera *et al.* (23) proposed a novel convolutional module as the first attempt to utilize the neighborhood similarity of

6530

Zhang et al. LDCT image denoising using deep attention network

CT images for denoising tasks. The query-specific GC modeling mechanism in an NLnet needs to generate huge attention maps to measure the relationships for each query position pair. Since the input features images always have high resolution in CT imaging tasks, NLnet-based methods have high computation complexity, which makes their integration into multiple layers problematic, resulting in ineffective modeling of the global contextual information in these layers.

Through a rigorous empirical analysis, Cao *et al.* (24) found that the GCs modeled with the NLnet are almost the same for different query positions within an image. Based on this finding, they created a simplified network based on a query-independent formulation, called the GC network, which maintains the accuracy of NLnet but with significantly less computation. The lightweight property of GC block allows it to be applied to multiple layers, leading to a better performance than that of the NLnet. The GC network aggregates the features of all positions together to form a GC feature for a feature image. Furthermore, different tissue structures and lesion changes generally vary greatly within a CT image, which leads to large statistical differences for the local neighbor regions containing distinct tissue structures or lesions. This, however, cannot be well described by a single GC feature as done in the GC network. This deviation in prior knowledge deviation from the real CT images limits the capability of such a useful GC modeling scheme and invites news development to further strengthen the field of CT image noise suppression. To this end, we propose an adaptive GC (AGC) modeling scheme for better representing the local contextual semantic information of CT images with much a lower computation cost than that of NLnet.

As for the network training, it is known that reducing the per-pixel loss as that as of mean-square error (MSE) between the network output and the ground truth alone tend to make output images oversmoothed and increase the image blur (25). The same effect can also be observed in traditional neural network-based CT image denoising methods (16). In this study, we propose a compound loss that combines the L1 loss [or called mean absolute error (MAE) loss], adversarial loss, and self-supervised, multiscale perceptual loss to practically solve the oversmoothing problem.

The work most similar to ours is that of Yang *et al.* and Li *et al.* (17,22), who also adopted a combination of adversarial loss and perceptual loss to produce sharper results. Our work differs from theirs in many important ways, and we would like to highlight some key points below.

(I) We propose an AGC modeling scheme to describe the nonlocal correlations and the regionally distinct statistics in CT images. The proposed AGC model, which contains soft split, aggregation, and replacement procedures, aggregates locally contextual semantic information adaptively for each regional neighborhood (referred to as patch in this paper). Furthermore, with a soft split and replacement strategy, the strong correlations among surrounding patches can be considered, leading to a better preservation of fine structural information such as tissue edges and textures represented by surrounding patches.

(II) We further propose an AGC-based long-short RED (AGC-LSRED) network for efficient LDCT image noise reduction. Specifically, an encoder-decoder structure with long skip connections is adopted as the backbone of the proposed denoising network. To better extract deeper semantic features, we propose to use stack of residual AGC attention blocks (RAGCBs) with short skip connection as the feature extractor in each layer. The long and short skip connections allow the valuable structural and positional information to be bypassed through these identity-based skip connections, which can ease the training of the deep denoising network.

(III) We propose a compound loss to better preserve the fine structures of the denoised results. In the compound loss, we adopt the L1 loss to encourage data fidelity for the generator network, the adversarial loss to measure the discrepancy between distributions of ground truth images and resulting images for producing more realistic images, and the self-supervised multiscale perceptual loss to measure the difference between image features in terms of both low-level semantic features and high-level semantic features. Our study demonstrated that the proposed network can achieve satisfactory results in preserving fine anatomical structures and suppressing noise in LDCT images.

Methods

AGC modeling scheme

The GC module

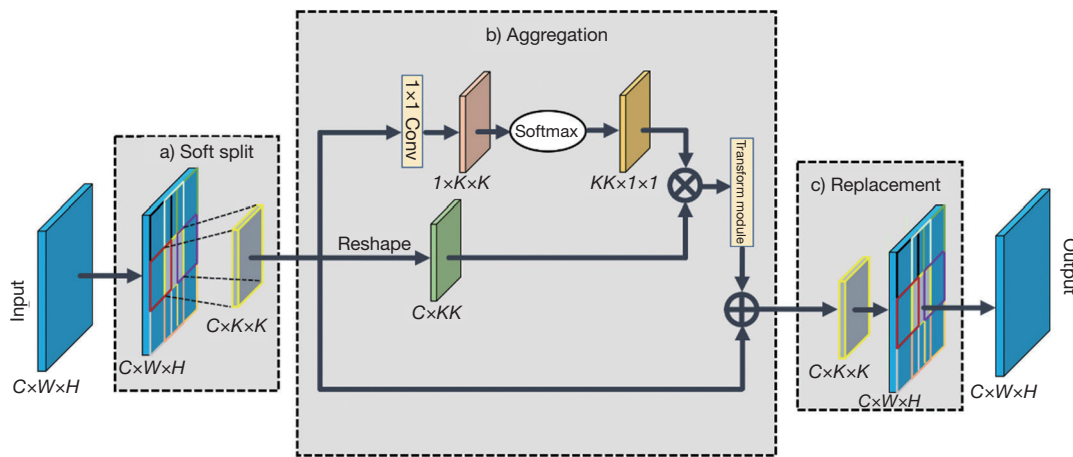The general GC modeling framework can be defined as

Figure 1 Architecture of the AGC module. The structure of the transform module is shown in *Figure 2*. 1×1 Conv, convolution layer with kernel size of 1×1; AGC, adaptive global context.
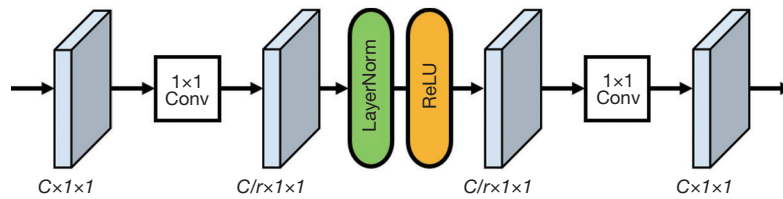

Figure 2 Structure of the transform module. 1×1 Conv, convolution layer with kernel size of 1×1; ReLU, rectified linear unit.

follows (24):

$$Z_i = X_i + \delta\left(\sum_{j=1}^{N_p} \alpha_j X_j\right) \qquad [1]$$

where $X \in \mathbb{R}^{C \times H \times W}$ and $Z \in \mathbb{R}^{C \times H \times W}$ denote the input and output feature image with a channel number of C, respectively; a height of H and width of W; i denotes the index of the query position; j enumerates all possible positions; $N_p = H \times W$ is the number of positions in the feature map; α_j is the aggregation weight; and $\delta(\cdot)$ is the feature transformation operation used to capture channel-wise dependencies which can be denoted as $\delta(\cdot) = W_{v2} ReLU(LN(W_{v1}(\cdot)))$, where W_{v1} and W_{v2} denote two linear transformations, respectively.

The proposed AGC module

On the basis of GC, we propose the AGC module to better describe the different local data statistics in CT feature image. The proposed AGC modeling mechanism consists of three processes: soft split, aggregation, and replacement, as illustrated in *Figures 1,2*.

(I) Soft split: we apply the soft split for modeling each

local contextual information. To avoid information loss, we split the CT image feature into patches with overlapping. For the input feature images $X \in \mathbb{R}^{C \times H \times W}$, suppose the size of each patch is $C \times k \times k$ with d overlapping, then the total of $L = \left\lfloor \frac{H-k}{k-d} + 1 \right\rfloor \times \left\lfloor \frac{W-k}{k-d} + 1 \right\rfloor$ patches can be extracted. After the soft split, the patches are input into the next process.

(II) Aggregation: we compute the GC information within a patch using the features of all positions within it and add the aggregated GC information to each query position of this patch to form the patch output. This process can be defined as follows:

$$Z_i^l = X_i^l + W_{v2} ReLU\left(LN\left(W_{v1}\left(\sum_{j=1}^{N_l} \alpha_j^l X_j^l\right)\right)\right),$$
$$\text{s.t.} \quad i = 1, \cdots, N_l, l = 1, \cdots, L \qquad [2]$$

where l denotes the index of the patch; N_l is the number of positions in the feature map of the lth patch; i denotes the index of the query position; and j enumerates all possible positions in it. For the weight α_j^l, we use the following Gaussian
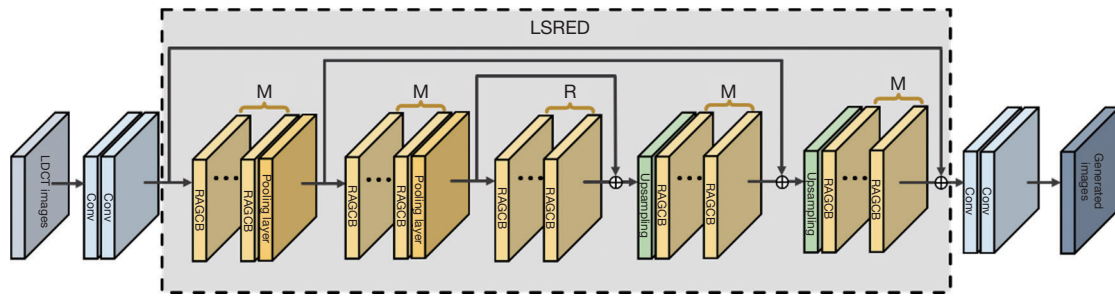
Figure 3 Structure of the AGC-LSRED generator network. ⊕, element-wise add operation. LDCT, low-dose computed tomography; conv, convolution layer; LSRED, long-short residual encoder-decoder; RAGCB, residual AGC attention block; AGC, adaptive global context; AGC-LSRED, AGC-based LSRED.
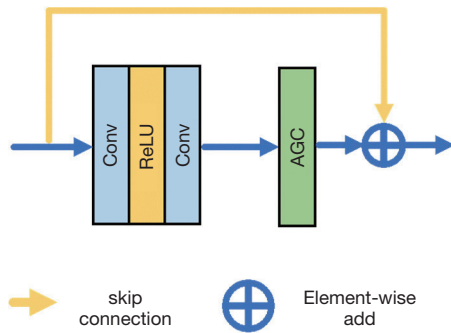

Figure 4 Structure of the proposed RAGCB module. Conv, convolution layer; ReLU, rectified linear unit; AGC, adaptive global context; RAGCB, residual AGC attention block.

embedding: $\alpha_j^i = \dfrac{e^{W_k X_j^i}}{\sum_{m=1}^{N_i} e^{W_k X_m^i}}$, where W_k is a linear transformation.

(III) Replacement: after the aggregation process, the GC-encoded patches are placed back to its position. For each position, there are multiple GC-encoded values from neighboring overlapping patches. We obtain the final value of each position by averaging its values from all patches overlapping it.

AGC-LSRED generator network

Network architecture

As illustrated in *Figure 3*, the proposed AGC-LSRED generator is mainly composed of three parts: shallow feature extraction, LSRED deep feature extraction, and feature refinement. Specifically, denoting the input LDCT image with X_{LD} and the output of the AGC-LSRED generator with Y_{LD}, we first use two consecutive convolution layers to extract the shallow features F_{SF} from the input X_{LD}. We

then use the proposed LSRED module to extract the deep features F_{DF} from F_{SF}. Finally, the extracted deep features are further refined with two consecutive convolution layers to form the final denoised output Y_{LD}. In the following sections, we describe the proposed LSRED deep feature extractor module in detail.

LSRED

Inspired by the work of Chen *et al.* (16), we use an encoder-decoder structure along with long skip connections as the backbone of the LSRED deep feature extractor, which contains RAGCBs with short skip connections, max-pooling downsamplings, bilinear interpolation upsamplings, and long skip connections, as shown in *Figure 3*. Specifically, in the encoder part of the LSRED, we first use two layers, which respectively contains *M* consecutive RAGCB modules followed with the max-pooling operation to extract the major deep structural features from the input shallow feature image F_{SF}, while discarding the detail structures. We then use *R* consecutive RAGCB modules to further refine these extracted features in a deeper embedding manifold and obtain the refined deep structural features. For the decoder part of the LSRED, two layers that respectively contain the upsampling operation and *M* consecutive RAGCB modules are adopted to reconstruct the deep textured structural features of the CT image, F_{DF}, from the information consolidated by the encoder. Three long skip connections are used to stabilize the train process.

The structure of the proposed RAGCB block is shown in *Figure 4*, which contains two stacked convolution layers, an AGC attention module, and a short skip connection. In each RAGCB module, the contextual semantic information within the feature images is adaptively captured by the proposed AGC modeling scheme. This kind of attention
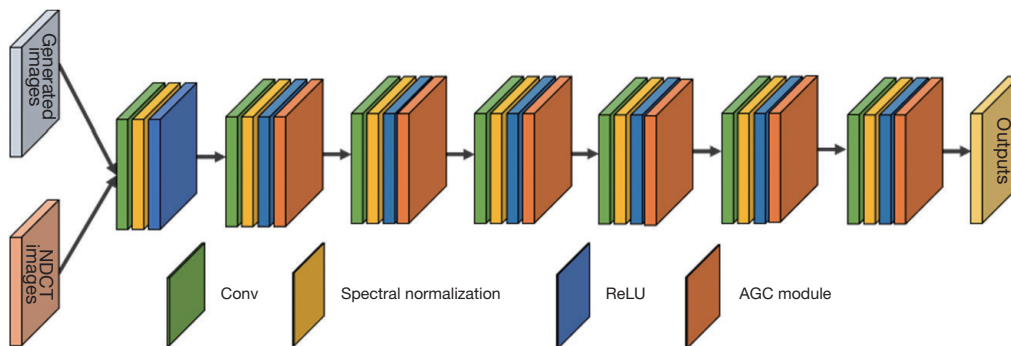
Figure 5 Structure of the AGC-SNMP discriminator network. NDCT, normal-dose computed tomography; conv, convolution layer; ReLU, rectified linear unit; AGC, adaptive global context; SNMP, spectral-normalized Markov patch; AGC-SNMP, AGC-based SNMP.

mechanism furnishes the proposed network with the ability to adaptively model the correlations among neighboring structures and hence enhance the representative learning ability.

In CT images, the preservation of subtle details and textures is highly desirable for clinical diagnosis, while the positional information is also critical for the localization of lesion changes. In the proposed LSRED module, the long and short skip connections can not only better guide the gradient backpropagation but also improve the information of the detailed structural and positional information from shallow layers to deep layers in a coarse level and a fine level, respectively, which helps the recovery of the underlying subtle details and textures for CT images.

AGC-attention-based discriminator network

The discriminator used in the proposed model is the same as that used in the method proposed by Bera *et al.* (23). The same spectral-normalized Markov patch (SNMP) discriminator is used as the backbone of the discriminator. The SNMP discriminator was first proposed in a patch-based general adversarial network (GAN) loss called spectral-normalized patch GAN (SN-PatchGAN) by Yu *et al.* (26). Compared with conventional discriminator, it can better focus on local locations and semantics. We further added our proposed AGC module to the SNMP discriminator network to adaptively capture the global contextual semantic information. The structure of the proposed AGC-based SNMP (AGC-SNMP) discriminator is shown in *Figure 5*.

Loss function

Adversarial loss

The adversarial loss encourages the generator to convert the data distribution from a high-noise version to a low-noise version. In this work, we adopt the Wasserstein distance as the adversarial loss, which is defined as follows:

$$\min_{G}\max_{D} L_{WGAN}(D,G) = \mathrm{E}_{Y \sim P_{Y_{ND}}}\left[D(Y)\right] - \mathrm{E}_{X \sim P_{X_{LD}}}\left[D(G(X))\right] + \lambda \mathrm{E}_{\hat{X}}\left[\left(\left\|\nabla_{\hat{X}} D(\hat{X})\right\|_2 - 1\right)^2\right] \quad [3]$$

where G and D are the proposed AGC-LSRED generator and AGC-SNMP discriminator, respectively; $P_{Y_{ND}}$ and $P_{X_{LD}}$ denote the distribution of the normal-dose ground truth CT images and noisy LDCT images, respectively; $\hat{X}$ is sampled uniformly along a straight line connecting pairs of generated samples and real samples; and λ is a weighting parameter. The generator G and the discriminator D are trained alternately by fixing one and updating the other.

L_1 loss

In this study, we used the L_1 loss to encourage data fidelity for the generator network. Compared with the L_2 loss (i.e., the MSE loss), the L_1 loss does not overpenalize large differences or tolerate small errors between the estimated image, leading to better preservation of details and textures. The L_1 loss is defined as follows:

$$L_{L_1}(G) = \mathrm{E}_{(X \sim P_{X_{LD}}, Y \sim P_{Y_{ND}})}\left[\left\|G(X) - Y\right\|_1\right] \quad [4]$$

Self-supervised multiscale perceptual loss

Perceptual loss, which is used to simulate human vision mechanism, compares the denoised image and the ground-
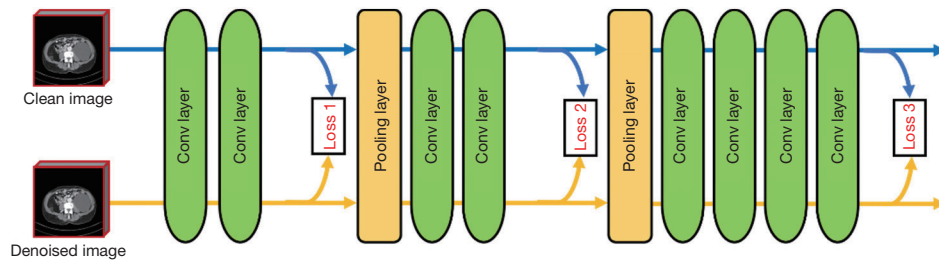
Figure 6 Architecture of self-supervised multiscale perceptual loss. Conv, convolution.

truth image in a feature manifold. Previous studies (17,27) have demonstrated that it can achieve improved results in terms of fine structure preservation. The visual geometry group (VGG) was been widely used as the feature extractor in previous works (17,28). Considering that the VGG feature extractor was originally trained for classifying natural images and thus might cause a loss of important domain-specific information for CT images. Li *et al.* (22) designed an autoencoder neural network and proposed a self-supervised learning scheme to train it. In this study, we adopted the same network structure and self-supervised learning strategy as that of Li *et al.* (22) to extract features for our perceptual loss design. In the perceptual loss network proposed by Li *et al.*, only the output features of the last layer of the encoder network are used for image feature comparison. In this study, we instead employed the output features of each layer of the encoder for the image feature comparison, as demonstrated in *Figure 6*. With such a multiscale perceptual loss, the generator has the ability to compare the denoised result against the ground truth image in terms of both low-level semantic features and high-level semantic features, thus leading to a better performance for preserving both major and subtle structures. The proposed self-supervised multiscale perceptual loss can be defined as follows:

$$L_{Perceptual}(G) = \mathrm{E}_{\left[X \sim P_{X_{LD}}, Y \sim P_{Y_{ND}}\right]}\left[\sum_{i=1}^{3}\left\|\phi_i\left(G(X)\right) - \phi_i\left(G(Y)\right)\right\|_F^2\right] \quad [5]$$

where ϕ_i denotes the ith feature extractor in the encoder.

The total loss for training our AGC-LSRED network can be expressed as follows:

$$L_{AGC\text{-}LSRED} = L_{L_1}(G) + \lambda_1 L_{WGAN}(D,G) + \lambda_2 L_{Perceptual}(G) \quad [6]$$

where λ_1 and λ_2 are 2 manual weighting parameters.

Datasets

In this work, the American Association of Physicists in Medicine (AAPM)-Mayo dataset was used to evaluate and

validate the proposed AGC-LSRED denoising method. This AAPM-Mayo dataset is a real clinical dataset licensed by Mayo Clinic for the 2016 National Institutes of Health (NIH)-AAPM-Mayo Clinic LDCT Grand Challenge (29). The dataset contains normal-dose abdominal CT images and quarter-dose CT images from 10 anonymous patients. In our experiments, we used CT images with a 3-mm slice thickness of 9 patients as the training set, comprising 4334 CT images, and we used CT images of 1 patient (L506) as the test set, comprising 422 CT images.

In addition, normal-dose CT and LDCT scans acquired from clinical CT colonography studies were used to further evaluate and validate the proposed AGC-LSRED method. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the institutional ethics committee of the Fourth Military Medical University. Informed consent was obtained from all the patients. The normal-dose scan was first acquired using an uCT760 CT scanner (United Imaging Healthcare, Brooklyn, NY, USA) at an X-ray tube voltage of 120 kVp and a tube current of 98 mAs. This was followed by low-dose scanning at an X-ray tube voltage of 100 kVp and a tube current of 24 mAs. The other scanning parameters were as follows: 0.579 s per gantry rotation, 3-mm slice thickness, and voxel size 0.7617×0.7617 mm². The reconstructed image was of 512×512 size.

Results

Parameter setting

In our experiments, we set the size of convolutional kernel of each convolutional layer in the generator and the discriminator to be 3×3 and the number of channels to be 64. In the AGC-LSRED generator network, we empirically set M=4 and R=6. In the soft split of the AGC module, we set k=15 and d=8 for the first layer of the encoder and the last layer of the decoder in the LSRED
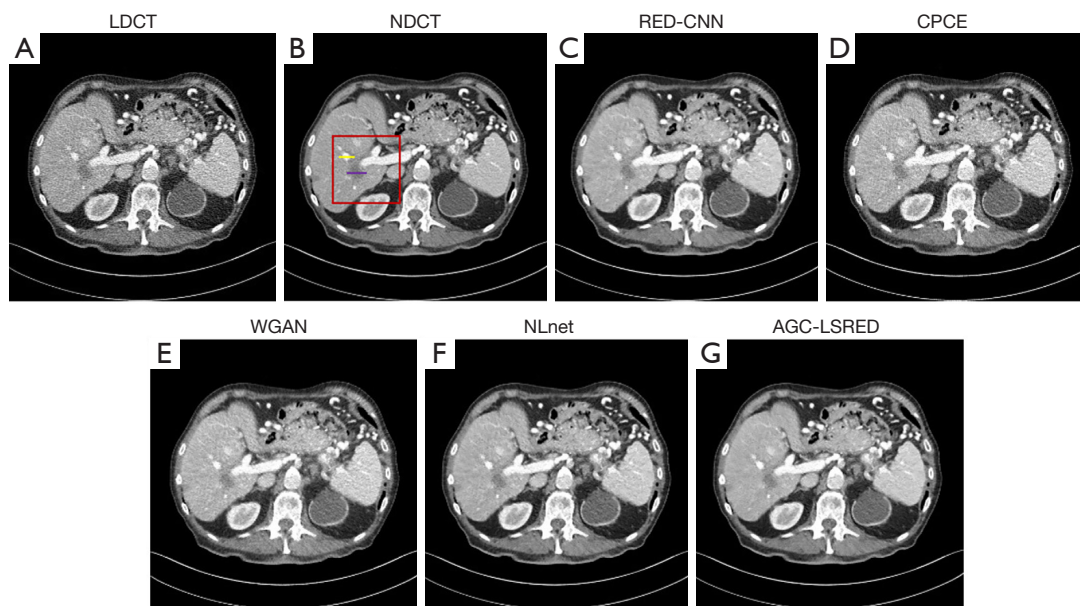
Figure 7 Processing results of the first representative slice of patient L506. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The yellow line indicates vessel and the purple line indicates the liver nodule. The red rectangle shows in *Figure 8*. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.

and we set k=8 and d=6 for the other layer of LSRED. In order to train the network, we took 20 randomly cropped 64×64 blocks from each slice, resulting in a total of 86,680 training blocks. The batch size was set to 8. We initially set the learning rate to be 1e–4 for the generator network and 4e–4 for the discriminator network. The two learning rates were both set to be decreased by a factor of two every 6,000 iterations. We set the parameters λ_1 and λ_2 to both be 0.1. For the Wasserstein GAN (WGAN) training, the weighting parameter λ that controls the tradeoff between Wasserstein distance and gradient penalty was set to 10. We used the Adam optimizer to train the network. We trained the network until the loss did not improve after 200 epochs. The networks were implemented using PyTorch and were trained/tested on an artificial intelligence (AI) workstation equipped with an Nvidia Tesla V100 GPU.

We compare the proposed method with the recently developed state-of-the-art deep learning-based LDCT denoising algorithms, including RED-CNN (16), conveying path-based convolutional encoder-decoder (CPCE) (28), WGAN (17), and NLnet (23). For these comparison methods, the AAPM-Mayo dataset was also used to train the network, and the network parameters were set as they

were described in their literatures. We also conduct an ablation study to demonstrate effects of the proposed AGC module and the self-supervised multiscale perceptual loss. The source codes are available online (https://github.com/Frank-ZhangYK/AGC-LSRED).

Experimental results of AAPM-Mayo data

Visual evaluation

To show the denoising effect of the proposed network, we selected the visualization results of three representative slices of test patient L506, as shown in *Figures 7-12*, where *Figures 8,10,12* show the zoomed region of interest (ROI) marked by red rectangles in *Figures 7,9,11*, respectively. The display windows of *Figures 7-12* are all set to –160,240 Hounsfield units (HU). It can be observed that all the deep learning-based methods can suppress the noise. Compared with other methods, the proposed AGC-LSRED method performs much better in terms of both noise suppression and fine structure preservation.

We further illustrate the coronal view of the test patient L506 in *Figures 13,14*. We can observe that the proposed AGC-LSRED method provides more homogeneous
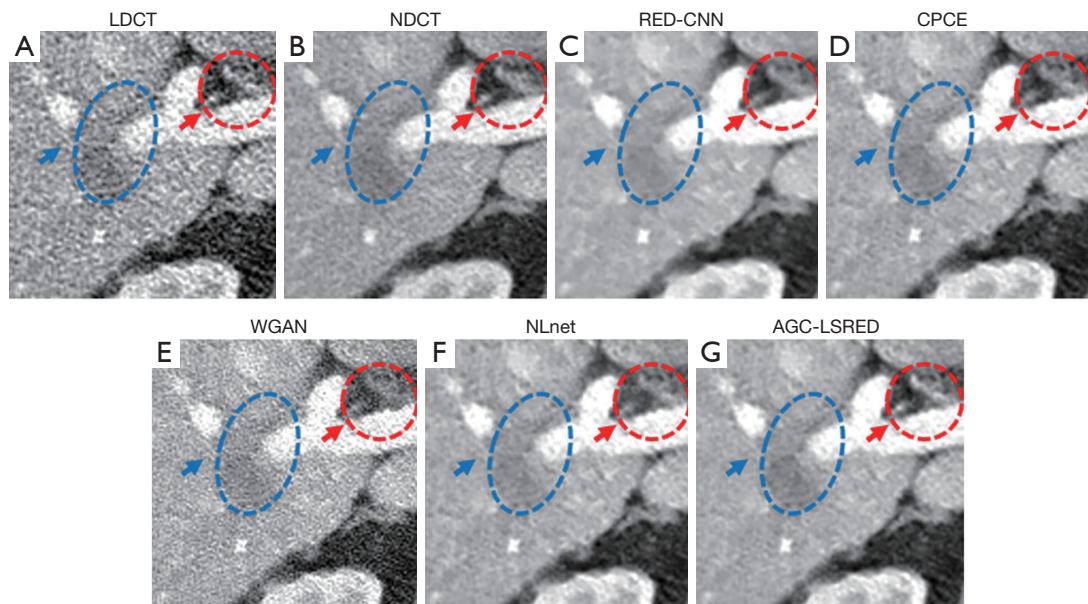
Figure 8 Zoomed images of a selected region, outlined by the red rectangle in *Figure 7B*. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The blue ellipses indicate the low-attenuation lesions in the posterior hepatic lobe, and the red ellipses indicate the tissue structures. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.
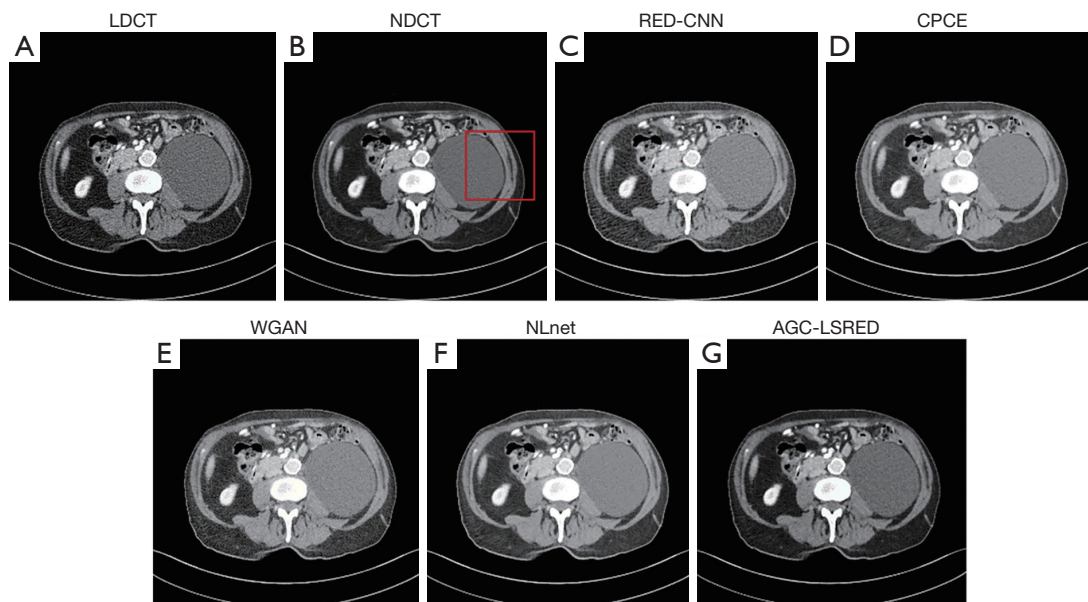

Figure 9 Processing results of the second representative slice of patient L506. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The red rectangle shows in *Figure 10*. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.
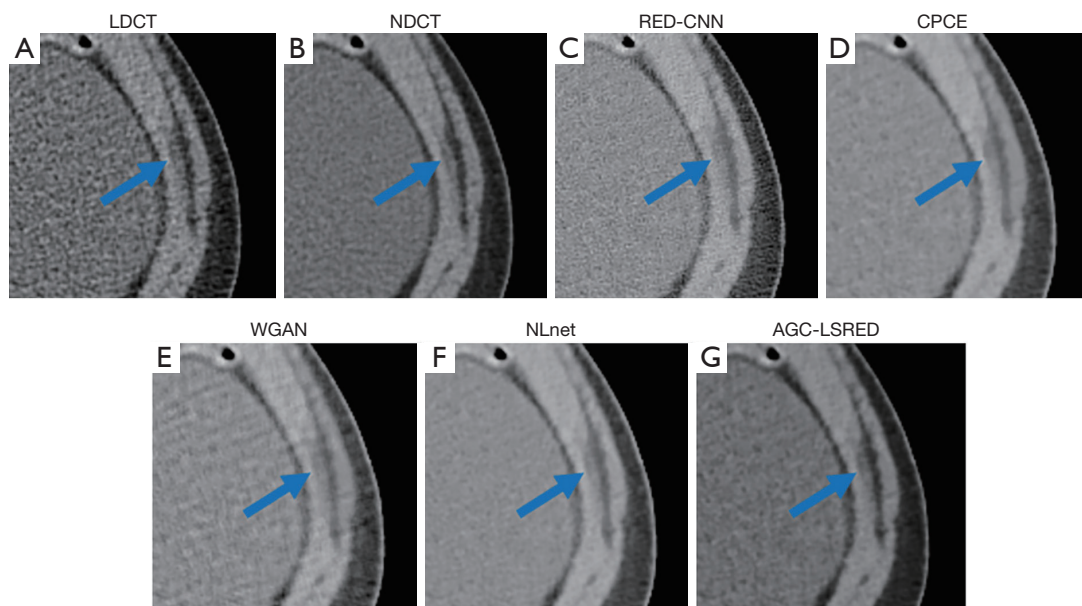
Figure 10 Zoomed images of a selected region, outlined by the red rectangle in *Figure 9B*. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The blue arrows indicate the obvious visual difference of the proposed method. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.
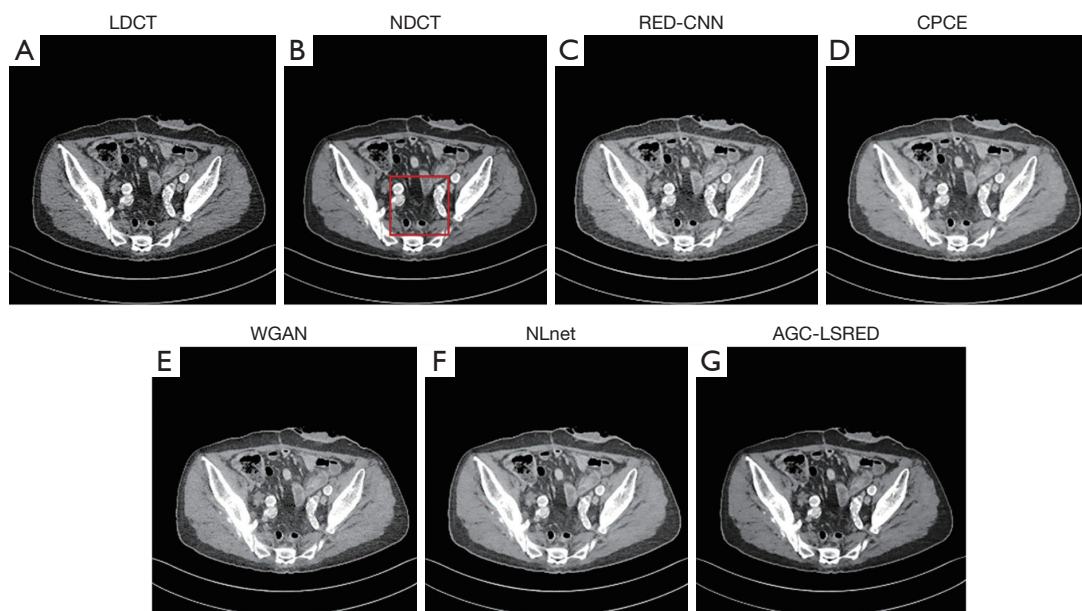

Figure 11 Processing results of the third representative slice of patient L506. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The red rectangle shows in *Figure 12*. LDCT, low-dose computed tomography; NDCT, normal-dose CT; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.
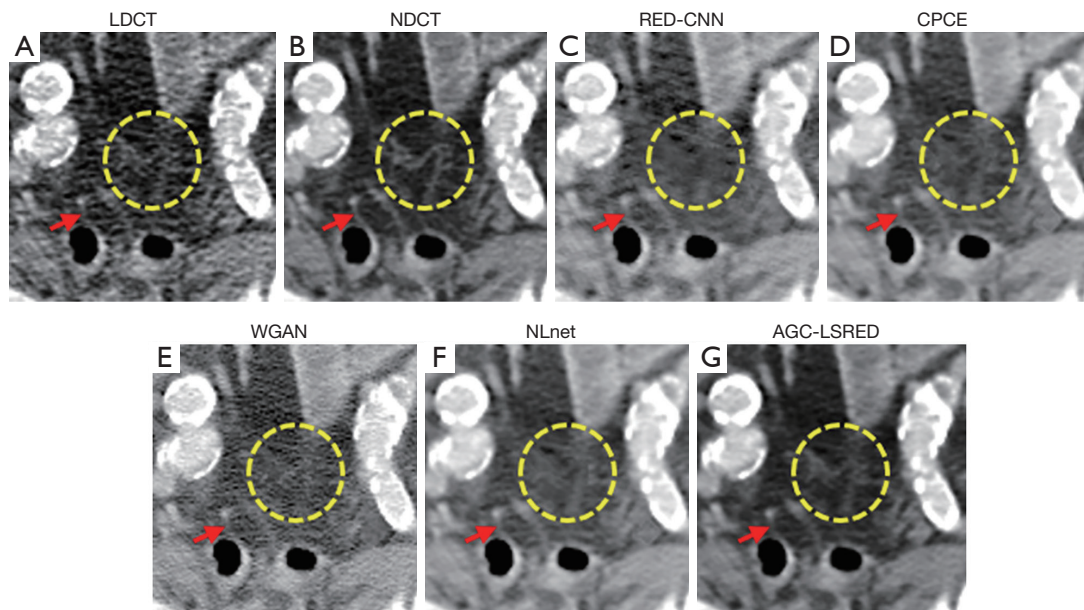
Figure 12 Zoomed images of a selected region outlined by the red rectangle in *Figure 11B*. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The red arrows and the yellow ellipses indicate the visual obvious difference of the proposed method. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.

processing results with better performance of fine structure preservation compared with other methods, especially for the selected ROI containing the suspected liver nodule lesion (as outlined by the red rectangle in *Figure 13B*).

To further compare the performance differences between RED-CNN, CPCE, WGAN, NLnet, and the proposed AGC-LSRED, we drew the intensity profiles through the vessel (along the yellow line in *Figure 7B*) and liver nodule (along the purple line in *Figure 7B*) in *Figure 15A,15B*, respectively. Compared with other methods, the results obtained by the proposed method are more consistent with the ground truth. The results demonstrate the proposed AGC-LSRED method performs better in preserving structures of the organ tissues.

Quantitative evaluation

To further illustrate the effectiveness of the proposed method, we quantitatively calculate the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), and the root-MSE (RMSE) values. *Table 1* summarizes the comparative results for each method. It demonstrates the proposed AGC-LSRED method exhibits the best result

with the lowest RMSE and the highest PSNR and SSIM.

Haralick texture measures

To further validate the effectiveness of the proposed AGC-LSRED method on texture preservation, Haralick texture feature measurement (30) was used in this study. Haralick texture features were extracted from the regions marked with the red rectangle in *Figure 7B*. The corresponding ROI of the normal-dose CT images was used as the baseline. We extracted 13 Haralick texture features from the ROIs and then calculated the normalized Euclidean distance between the features of the reference image and the processed results. The normalized Euclidean distances were then calculated for the reference image and the images were processed using the RED-CNN, CPCE, WGAN, NLnet, and proposed method. A shorter distance indicates better texture preservation. The corresponding results are shown in *Table 2*. The gain of our proposed method in preserving the abdominal tissue texture is obvious.

Ablation analysis

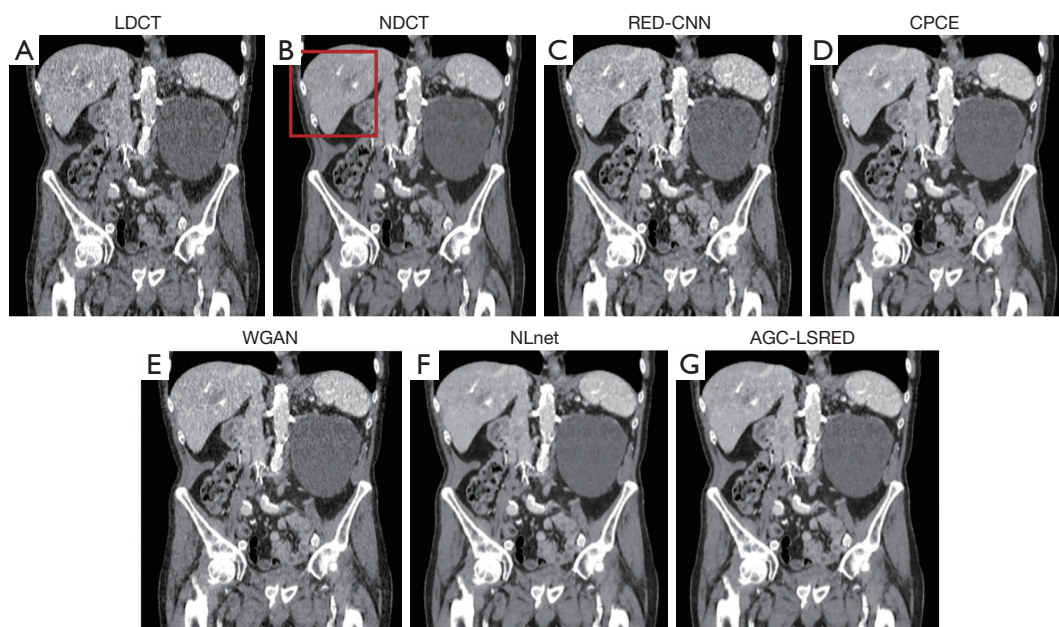We completed an ablation study to identify effects of the

Figure 13 Coronal section view of case L506 from the AAPM-Mayo LDCT dataset. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The red rectangle shows in *Figure 14*. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; AAPM, American Association of Physicists in Medicine; HU, Hounsfield units.
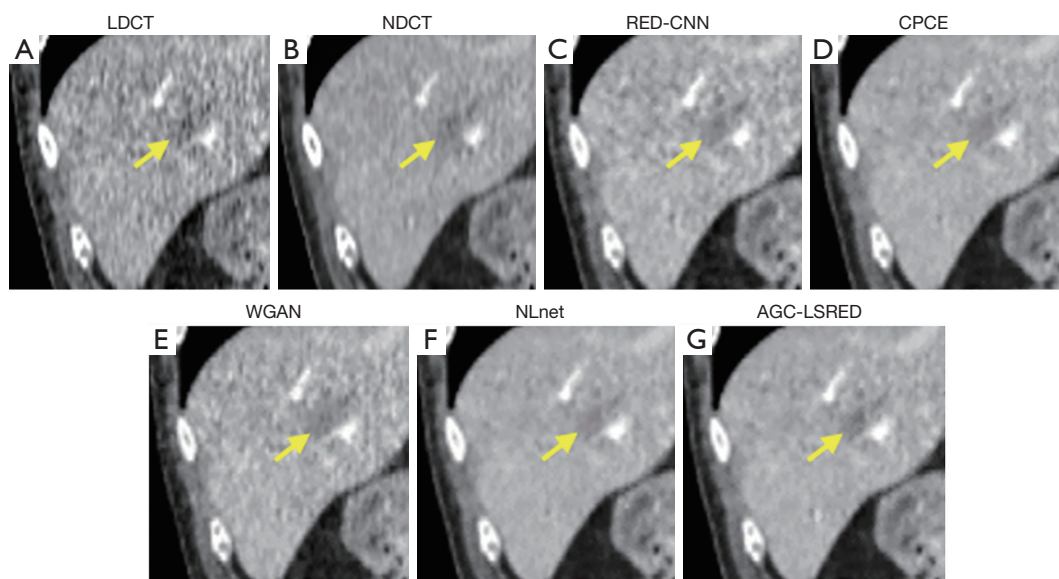

Figure 14 Zoomed images of a selected region outlined by the red rectangle in *Figure 13B*. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The yellow arrows indicate the obvious visual difference of the proposed method. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.
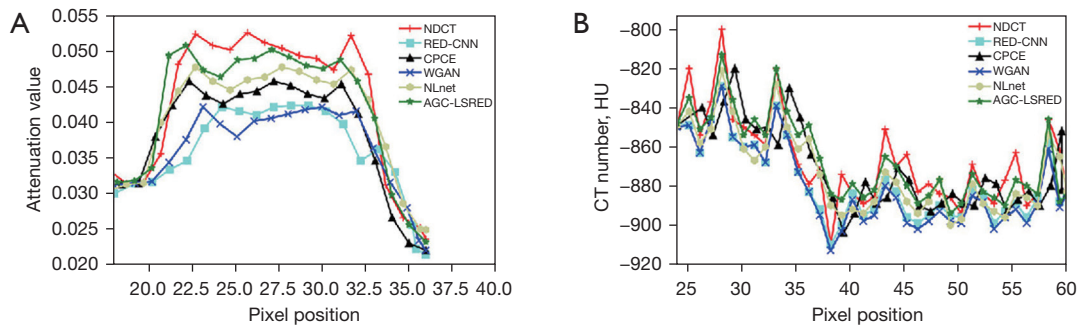
Figure 15 Intensity proles along the horizontal lines labeled in *Figure 7B*. (A) Intensity proles along the yellow line. (B) Intensity proles along the purple line. NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder.

Table 1 Averaged RMSE, PSNR, and SSIM values from the processed results of all the test slices

Methods	LDCT	RED-CNN	CPCE	WGAN	NLnet	AGC-LSRED
RMSE	14.24	9.28	9.25	10.77	9.11	9.02[†]
PSNR	27.24	32.93	33.04	30.80	33.06	33.17[†]
SSIM	0.853	0.910	0.913	0.893	0.916	0.925[†]

[†], the best results. RMSE, root-mean-square error; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index; LDCT, low-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder.

Table 2 Normalized Haralick texture distance between the ROI (as indicated in *Figure 7B*) of the reference image and that of the reconstructed results

Tissue type	RED-CNN	CPCE	WGAN	NLnet	AGC-LSRED
Abdomen (ROI I)	0.0058	0.0054	0.0062	0.0046	0.0023[†]

[†], the best results. ROI, region of interest; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder.

Table 3 Summary of all trained networks: loss functions and trainable networks

Variant name	Comment
C1	Network with GC modules trained with self-supervised multiscale perceptual loss
C2	Network with AGC modules trained with self-supervised single-scale perceptual loss
C3	Network with AGC modules trained with self-supervised multiscale perceptual loss

GC, global context; AGC, adaptive global context.

proposed AGC module and the self-supervised multiscale perceptual loss. To this end, we considered three variations of the proposed AGC-LSRED network for comparison, as shown in *Table 3*.

Effectiveness of the AGC module

First, performed a comparison between the AGC module and the GC module. The quantitative values of the processing results using C1 and C3 are shown in *Table 4*.

Table 4 Quantitative results for the relevant metrics of the test dataset

Methods	C1	C2	C3 (ours)
RMSE	9.11	9.09	9.02[†]
PSNR	33.08	33.12	33.17[†]
SSIM	0.918	0.921	0.925[†]

[†], the best results. RMSE, root-mean-square error; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index.

It was found that the C3 method (with the AGC module) performs better than does the C1 method (with the GC module).

Effectiveness of the self-supervised multiscale perceptual loss

In terms of the denoising performance of the perceptual loss function, we compared the C3 method (with self-supervised multiscale perceptual loss) with the C2 method (with self-supervised single-scale perceptual loss). The quantitative results are shown in *Table 4*. The quantitative results demonstrate that using multiscale perceptual loss provides a better performance than does using single-scale perceptual loss, which verifies the effectiveness of self-supervised multiscale perceptual loss.

Experimental results of clinical patient data

Visual evaluation

In this pilot clinical study, the 100 kVp/24 mAs LDCT scan from a patient was used for the evaluation, as shown in *Figure 16A*. The corresponding 120 kVp/98 mAs normal-dose scans from the same patient were used as the reference images, as shown in *Figure 16B*. *Figures 16,17* show the resulting images. When using real clinical data, the proposed method produces a most similar visual effect to the normal-dose reference scan and performs better than do other methods with respect to noise suppression and structure preservation.

Evaluation by radiologists

A total of 63 slices of the 100 kVp/24 mAs low-dose scan were independently scored by three radiologists in terms of noise reduction and structure and texture preservation. All the images to be evaluated were randomly displayed on the screen. The score ranged from 0 (worst) to 5 (best). The average scores of each radiologist for each image subset are presented in *Table 5*. The proposed AGC-LSRED algorithm demonstrated advantages over other methods in terms of subjective assessment scores.

Discussion

This paper proposes an AGC modeling scheme to characterize the nonlocal correlations and the regionally distinct statistics in CT images. The proposed AGC modeling mechanism contains three processes, which are soft split, aggregation, and replacement. In this manner, the locally contextual semantic information can be aggregated adaptively for each regional neighborhood. In addition, the strong correlations among surrounding patches can be considered with the soft split and replacement strategy, which helps to better preserve the fine structural information such as tissue edges and textures represented by the surrounding patches.

Various attention networks (31-34) have been developed in the past few years. In this study, the proposed AGC was developed on the basis of the GC attention modeling scheme. We opted for the GC-based modeling scheme (24) mainly because it can effectively model the GC as do NLnet and dual attention network (DANet) (31) (which is a heavy weight and difficult to integrate into multiple layers) with the lightweight property as do squeeze-and-excitation network (SENet) (32), convolutional block attention module network (CBAM-Net) (33), and residual attention network (34) (which adopts rescaling for feature fusion and is not sufficiently effective for GC modeling). Combining the channel attention, as is done in DANet (31) and global second-order pooling convolutional network (GSoP-Net) (35), can be expected to improve the LDCT image denoising performance, and in our future work, we intend to investigate this possibility further. More recently, vision transformer (36), a full self-attention mechanism, originally designed for natural language processing (NLP) (37), has shown the state-of-the-art performance in several vision problems, including image classification (36), object detection (38), and image restoration (39). In future work, we aim to combine the proposed LSRED network framework with the vision transformer modeling scheme so as to better capture global interactions between contexts. Further improvement in noise suppression and fine structure preservation for LDCT images is expected.
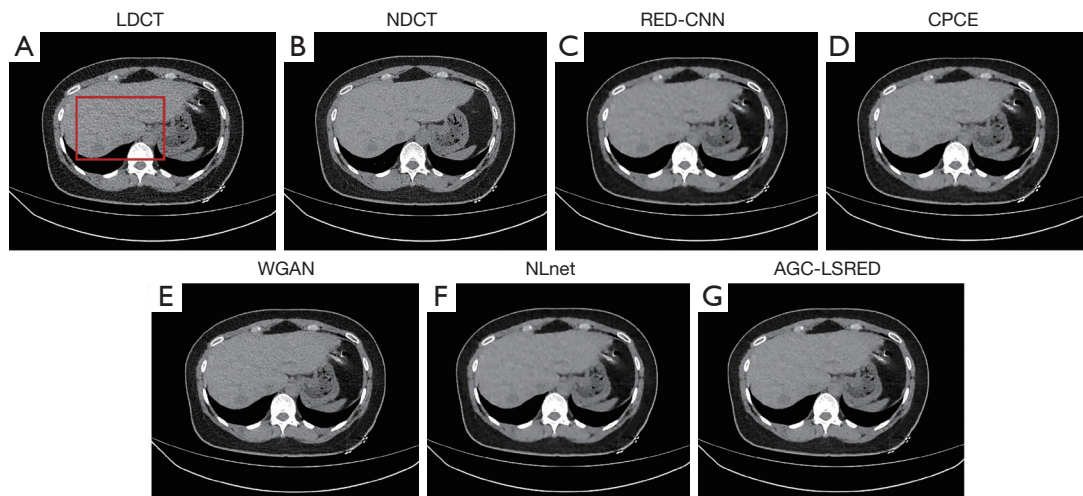
Figure 16 Processing results of 1 slice of the real 100 kVp/24 mAs scan. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. The red rectangle shows in *Figure 17*. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; mAs, milliampere-seconds; HU, Hounsfield units.
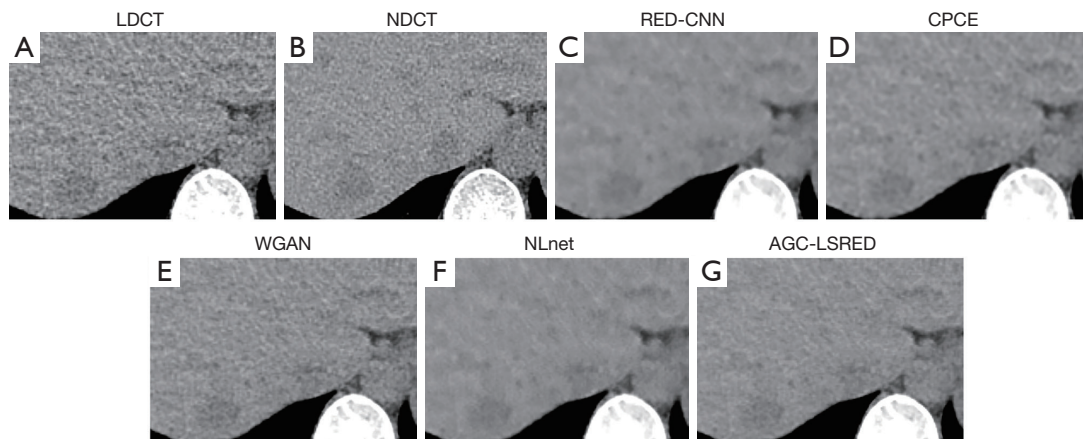

Figure 17 Zoomed images of a selected region outlined by the red rectangle in *Figure 16A*. (A) LDCT; (B) NDCT; (C) RED-CNN; (D) CPCE; (E) WGAN; (F) NLnet; (G) AGC-LSRED. The display window is –160,240 HU. LDCT, low-dose computed tomography; NDCT, normal-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder; HU, Hounsfield units.

Conclusions

We propose AGC-LSRED network to improve the performance of the structure-preserving LDCT image noise reduction task. The backbone of the proposed AGC-LSRED network is an encoder-decoder structure with long skip connections. For each layer, we use the stack of residual AGC-attention blocks with short skip connection as the feature extractor. The proposed denoising model can benefit the information flow of the structural semantic information from shallow layers to deep layers in a coarse level and a fine level, respectively, thus helping the recovery of the underlying subtle details and textures for CT images.

Table 5 Radiologists' scoring of the image quality

Radiologist	LDCT	RED-CNN	CPCE	WGAN	NLnet	AGC-LSRED
Radiologist #1	3.26	4.12	4.23	4.09	4.18	4.32[†]
Radiologist #2	3.52	4.15	4.31	4.17	4.20	4.46[†]
Radiologist #3	3.03	4.08	4.19	4.05	4.16	4.25[†]
Averaged scores	3.27	4.12	4.24	4.10	4.18	4.34[†]

[†], the best results. LDCT, low-dose computed tomography; RED-CNN, residual encoder-decoder convolutional neural network; CPCE, conveying path-based convolutional encoder-decoder; WGAN, Wasserstein general adversarial network; NLnet, nonlocal network; AGC-LSRED, adaptive global context-based long-short residual encoder-decoder.

To train the proposed AGC-LSRED network, we propose a compound loss that combines the L1 loss, adversarial loss, and perceptual loss for better preserving the fine structures of the denoised results. Compared with conventional perceptual loss, the proposed self-supervised multiscale perceptual loss provides the generator with the ability to compare the denoised result against the ground-truth image in terms of both low-level semantic features and high-level semantic features, thus leading to a better performance in preserving of both major and subtle structures.

LDCT data from the AAPM-Mayo clinical dataset and real clinical CT colonography studies were used to evaluate the proposed AGC-LSRED denoising method. The results indicate that the proposed method is superior for both noise suppression and fine structure preservation compared with the other competitive CNN-based methods.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-194/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the institutional ethics committee of the Fourth Military Medical University. Informed consent was obtained from all the patients.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Fu BJ, Lv ZM, Lv FJ, Li WJ, Lin RY, Chu ZG. Sensitivity and specificity of computed tomography hypodense sign when differentiating pulmonary inflammatory and malignant mass-like lesions. Quant Imaging Med Surg 2022;12:4435-47.

2. The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103. Ann ICRP 2007;37:1-332.

3. Keller G, Hagen F, Neubauer L, Rachunek K, Springer F, Kraus MS. Ultra-low dose CT for scaphoid fracture detection-a simulational approach to quantify the capability of radiation exposure reduction without diagnostic limitation. Quant Imaging Med Surg 2022;12:4622-32.

4. Li T, Li X, Wang J, Wen J, Lu H, Hsieh J, Liang Z. Nonlinear sinogram smoothing for low-dose X-ray CT. IEEE Trans Nucl Sci 2004;51:2505-13.

5. Zhang Y, Zhang J, Lu H. Statistical sinogram smoothing for low-dose CT with segmentation-based adaptive

6544

Zhang et al. LDCT image denoising using deep attention network

filtering. IEEE Trans Nucl Sci 2010;57:2587-98.

6. Hara AK, Paden RG, Silva AC, Kujak JL, Lawder HJ, Pavlicek W. Iterative reconstruction technique for reducing body radiation dose at CT: feasibility study. AJR Am J Roentgenol 2009;193:764-71.

7. Zhang Y, Peng J, Zeng D, Xie Q, Li S, Bian Z, Wang Y, Zhang Y, Zhao Q, Zhang H, Liang Z, Lu H, Meng D, Ma J. Contrast-Medium Anisotropy-Aware Tensor Total Variation Model for Robust Cerebral Perfusion CT Reconstruction with Low-Dose Scans. IEEE Trans Comput Imaging 2020;6:1375-88.

8. Ma J, Huang J, Feng Q, Zhang H, Lu H, Liang Z, Chen W. Low-dose computed tomography image restoration using previous normal-dose scan. Med Phys 2011;38:5713-31.

9. Zhang Y, Rong J, Lu H, Xing Y, Meng J. Low-Dose Lung CT Image Restoration Using Adaptive Prior Features From Full-Dose Training Database. IEEE Trans Med Imaging 2017;36:2510-23.

10. Li Z, Yu L, Trzasko JD, Lake DS, Blezek DJ, Fletcher JG, McCollough CH, Manduca A. Adaptive nonlocal means filtering based on local noise level for CT denoising. Med Phys 2014;41:011908.

11. Zhang Y, Lu H, Rong J, Meng J, Shang J, Ren P, Zhang J. Adaptive non-local means on local principle neighborhood for noise/artifacts reduction in low-dose CT images. Med Phys 2017;44:e230-41.

12. Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 2006;54:4311-22.

13. Sheng K, Gou S, Wu J, Qi SX. Denoised and texture enhanced MVCT to improve soft tissue conspicuity. Med Phys 2014;41:101916.

14. Zhang Y, Zeng D, Bian Z, Lu H, Ma J. Weighted tensor low-rankness and learnable analysis sparse representation model for texture preserving low-dose CT reconstruction. IEEE Trans Comput Imaging 2021;7:321-36.

15. Mendrik AM, Vonken EJ, Rutten A, Viergever MA, van Ginneken B. Noise reduction in computed tomography scans using 3-d anisotropic hybrid diffusion with continuous switch. IEEE Trans Med Imaging 2009;28:1585-94.

16. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. IEEE Trans Med Imaging 2017;36:2524-35.

17. Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G. Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. IEEE Trans Med Imaging 2018;37:1348-57.

18. Zavala-Mondragon LA, Rongen P, Bescos JO, de With PHN, van der Sommen F. Noise Reduction in CT Using Learned Wavelet-Frame Shrinkage Networks. IEEE Trans Med Imaging 2022;41:2048-66.

19. Zhang Y, Zhang W, Lei Y, Zhou J. Few-view image reconstruction with fractional-order total variation. J Opt Soc Am A Opt Image Sci Vis 2014;31:981-95.

20. Chen Y, Gao D, Nie C, Luo L, Chen W, Yin X, Lin Y. Bayesian statistical reconstruction for low-dose X-ray computed tomography using an adaptive-weighting nonlocal prior. Comput Med Imaging Graph 2009;33:495-500.

21. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018:7794-803.

22. Li M, Hsu W, Xie X, Cong J, Gao W. SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network. IEEE Trans Med Imaging 2020;39:2289-301.

23. Bera S, Biswas PK. Noise Conscious Training of Non Local Neural Network Powered by Self Attentive Spectral Normalized Markovian Patch GAN for Low Dose CT Denoising. IEEE Trans Med Imaging 2021;40:3663-73.

24. Cao Y, Xu J, Lin S, Wei F, Hu H. GCNET: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019:1971-80.

25. Huang R, Zhang S, Li T, He R. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: Proceedings of the IEEE International Conference on Computer Vision; 2017:2439-48.

26. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019:4471-80.

27. Johnson J, Alahi A, Li FF. Perceptual losses for real-time style transfer and super-resolution. In: Leibe B, Matas J, Sebe N, Welling M. editors. Computer Vision-ECCV 2016. Cham: Springer; 2016:694-711.

28. Shan H, Zhang Y, Yang Q, Kruger U, Kalra MK, Sun L, Cong W, Wang G. 3-D Convolutional Encoder-Decoder Network for Low-Dose CT via Transfer Learning From a 2-D Trained Network. IEEE Trans Med Imaging 2018;37:1522-34.

29. Low Dose CT Grand Challenge. Available online: https://www.aapm.org/GrandChallenge/LowDoseCT/

30. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;(6):610-21.

31. Fu J, Liu J, Tian H, Li Y. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019:3146-54.

32. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018:7132-41.

33. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018:3-19.

34. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017:3156-64.

35. Gao Z, Xie J, Wang Q, Li P. Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019:3024-33.

36. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [Preprint]. 2020. Available online: https://arxiv.org/abs/2010.11929

37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017:5998-6008.

38. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv:2010.04159 [Preprint]. 2020. Available online: https://arxiv.org/abs/2010.04159

39. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W. Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021:12299-310.