Original Article

Attention-based dual-branch deep network for sparse-view computed tomography image reconstruction

Xiang Gao[1,2], Ting Su[1], Yunxin Zhang[3], Jiongtao Zhu[1,4], Yuhang Tan[1], Han Cui[1], Xiaojing Long[1], Hairong Zheng[5], Dong Liang[1,5], Yongshuai Ge[1,5]

[1]Research Center for Medical Artificial Intelligence, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; [2]University of Chinese Academy of Sciences, Beijing, China; [3]Department of Vascular Surgery, Beijing Jishuitan Hospital, Beijing, China; [4]College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, China; [5]Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Contributions: (I) Conception and design: X Gao, T Su, Y Ge; (II) Administrative support: H Zheng, D Liang, Y Ge; (III) Provision of study materials or patients: Y Zhang, X Long, Y Ge; (IV) Collection and assembly of data: X Gao, T Su, Y Ge; (V) Data analysis and interpretation: X Gao, T Su, J Zhu, Y Tan, H Cui, Y Ge; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Ting Su. Research Center for Medical Artificial Intelligence, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Nanshan District, Shenzhen 518055, China. Email: ting.su@siat.ac.cn; Yongshuai Ge. Research Center for Medical Artificial Intelligence, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Nanshan District, Shenzhen 518055, China. Email: ys.ge@siat.ac.cn.

Background: The widespread application of X-ray computed tomography (CT) imaging in medical screening makes radiation safety a major concern for public health. Sparse-view CT is a promising solution to reduce the radiation dose. However, the reconstructed CT images obtained using sparse-view CT may suffer severe streaking artifacts and structural information loss.

Methods: In this study, a novel attention-based dual-branch network (ADB-Net) is proposed to solve the ill-posed problem of sparse-view CT image reconstruction. In this network, downsampled sinogram input is processed through 2 parallel branches (CT branch and signogram branch) of the ADB-Net to independently extract the distinct, high-level feature maps. These feature maps are fused in a specified attention module from 3 perspectives (channel, plane, and spatial) to allow complementary optimizations that can mitigate the streaking artifacts and the structure loss in sparse-view CT imaging.

Results: Numerical simulations, an anthropomorphic thorax phantom, and in vivo preclinical experiments were conducted to verify the sparse-view CT imaging performance of the ADB-Net. The proposed network achieved a root-mean-square error (RMSE) of 20.6160, a structural similarity (SSIM) of 0.9257, and a peak signal-to-noise ratio (PSNR) of 38.8246 on numerical data. The visualization results demonstrate that this newly developed network can consistently remove the streaking artifacts while maintaining the fine structures.

Conclusions: The proposed attention-based dual-branch deep network, ADB-Net, provides a promising alternative to reconstruct high-quality sparse-view CT images for low-dose CT imaging.

Keywords: CT reconstruction; sparse-view CT; streak artifact; deep learning; attention

Submitted Jun 15, 2022. Accepted for publication Dec 01, 2022. Published online Feb 10, 2023.
doi: 10.21037/qims-22-609
View this article at: https://dx.doi.org/10.21037/qims-22-609

Introduction

Over the past half century, computed tomography (CT) has become irreplaceable in modern medical imaging applications. However, concerns about the risks of radiation exposure have attracted considerable attention. According to the International Commission on Radiological Protection, every increase of 1 mSv in radiation dose in the human body increases the chance of canceration by nearly 1/20,000 (1). There is worldwide consensus to reduce the radiation dose of CT scans to as low as reasonably achievable. Dedicated research from the academic and industrial fields continues to advance low-dose CT imaging solutions.

Sparse-view CT scanning is a promising approach to achieving low-dose CT imaging. By downsampling the total number of acquired projections, the radiation dose received by patients could be dramatically reduced. In the low–milliampere-seconds scanning method, the tube output is reduced, and the noise of the CT images is substantially elevated (2-4). However, sparse-view CT images reconstructed from the conventional filtered back-projection (FBP) algorithm experience strong streaking artifacts and a loss of anatomical structure due to angular undersampling. Therefore, low-dose CT applications can be advanced by improving image quality in sparse-view CT imaging.

Over the past two decades, model-based iterative sparse-view CT image reconstruction methods have been investigated by reformulating the reconstruction task as a compression-aware optimization problem. In this approach, certain prior information and mathematical optimization models are assumed to jointly remove the streaks. However, because of the requirements of accurate forward model and parameter selections, the iterative CT reconstruction algorithms may have limitations in generating high-quality CT images (5-7). They may also have the drawback of a long running time in certain applications.

Recently, the development of deep learning techniques has opened many opportunities in medical image reconstruction fields and demonstrated excellent performance in CT reconstruction. For example, Jin *et al.* (8) applied the U-Net, and Han *et al.* (9) proposed the dual-frame U-Net via deep convolutional framelets (Frame U-Net) to perform sparse-view CT image reconstruction. Zhang *et al.* (10) proposed the DenseNet (11) and deconvolution-based network (DDNet) to join the advantages of the DenseNet and deconvolution operation to achieve promising results in CT reconstruction at a fast training speed. Kang *et al.* (12,13) combined the wavelet transform, residual block (14), and convolutional neural network (CNN) to remove the streaks. Li *et al.* (15) proposed the self-attention CNN (SACNN) to complete the self-supervised denoising process on low-dose CT images using the attention module.

Similarly, sinogram-based networks can also be used. For example, the sparse-view sinogram has been interpolated into a full-view sinogram via the U-Net in a deep neural network-enabled sinogram synthesis method (SS-Net) (16). Li *et al.* (17) proposed the iCT-net to convert a sparse-view sinogram directly into a high-quality CT image. Fu *et al.* (18) decomposed the sparse-view inverse problem into a set of simple transformations and used a layered network to perform CT reconstructions.

High-quality CT images can also be reconstructed using the information in both domains. For example, Lin *et al.* (19) applied the CT image and sinogram domain CNN to remove metal artifacts. Both Wang *et al.* (20) and Sun *et al.* (21) used an attention mechanism and the concept of dual domain to reconstruct the sparse-view CT images. One drawback of these approaches is the serial use of the 2 domains. The serial structure enables the 2-feature extractions to be computed sequentially, which may lead to incomplete feature extractions and inadequate compensation of the information lost in sparse-view CT imaging.

In this work, an innovative dual-branch end-to-end deep network based on the attention mechanism, attention-based dual-branch network (ADB-Net), is proposed to reconstruct high-quality sparse-view CT images directly from the downsampled sinogram (22-24). In ADB-Net, a unique network structure with 2 parallel feature extraction branches is designed. For the CT branch, the U-Net is used to extract the features of the CT images reconstructed from the conventional FBP algorithm. For the sinogram branch, convolutional layers and atrous spatial pyramid pooling (ASPP) are used to extract the global high-level features of the sinogram (25). Afterward, the attention mechanism is used to fuse the above independent feature maps to complement further feature extractions between the 2 network branches. Eventually, high-quality CT images with greatly mitigated streaking artifacts and structure loss are generated.

Methods

Problem formulation

In this sparse-view CT image reconstruction approach,

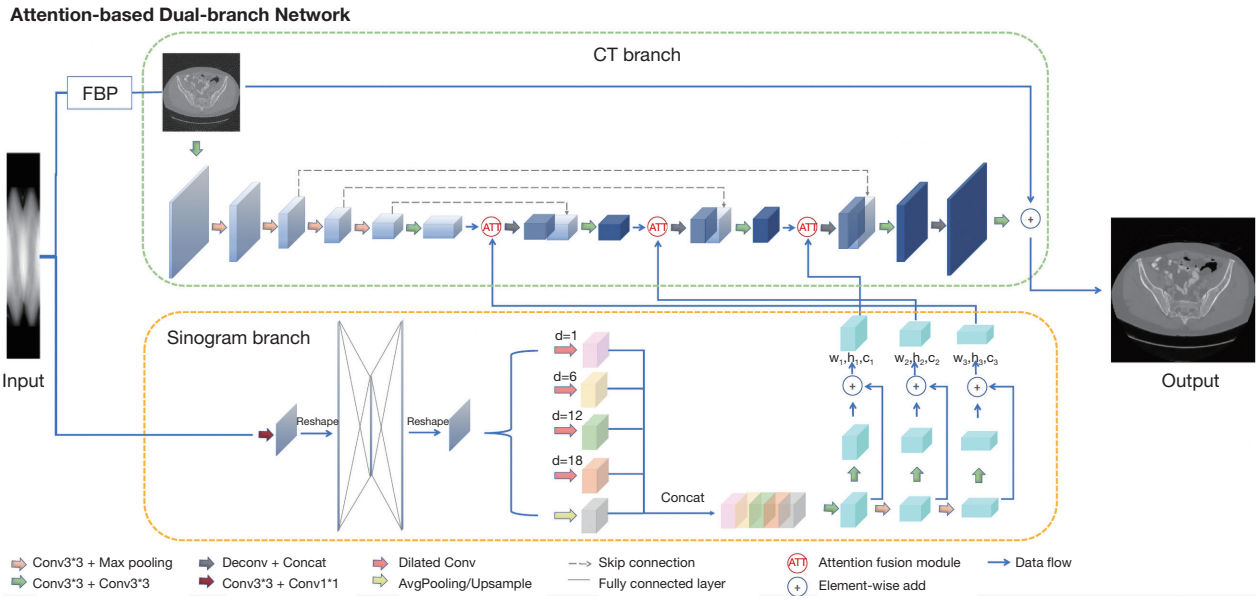

Figure 1 The architecture of the ADB-Net. The green box depicts the structure of the CT branch that accepts the FBP output, and the yellow box depicts the structure of the sino branch that accepts the sinogram input. Note that only the CT branch module and the sinogram branch module are depicted. The multiscale features that are extracted from the 2 branches are fed into the attention fusion module (represented as red circles). ADB-Net, attention-based dual-branch network; FBP, filtered back-projection; CT, computed tomography.

the network input is the sinogram $S \in \mathbb{R}^{W_{proj}*H_{proj}*C_{proj}}$, and the output is the high-quality CT image $Q \in \mathbb{R}^{W_{img}*H_{img}*C_{img}}$, with mitigated streaking artifacts. The $g(\cdot)$ represents the process of the network:

$$Q = g(F,S) = OP_{decoder}\left(OP_{fusion}\left(OP_F(F), OP_S(S)\right)\right) \quad [1]$$

where OP_F denotes the feature extraction operator for the CT image $F \in \mathbb{R}^{W_{img}*H_{img}*C_{img}}$ in the upper branch, OP_S denotes the feature extraction operator for the sinogram S in the lower branch, OP_{fusion} denotes the feature fusion operator implemented via the attention mechanism, and $OP_{decoder}$ denotes the feature decoding operator. The overall structure of $g(\cdot)$ is shown in *Figure 1*. The final network output is a residual image, which is added back to the FBP-reconstructed CT image to reduce the sparse-view streaking artifacts. The attention-based feature fusion module (i.e., OP_{fusion}) is illustrated in *Figure 2*.

CT branch

This component corresponds to the OP_F operator in Eq. [1]. In this branch, multiple high-level features of different scales can extracted, which makes the U-Net an attractive candidate to use for this takes (26). Feature maps with certain sizes (represented by $\{(w_1,h_1,c_1),(w_2,h_2,c_2),(w_3,h_3,c_3)\}$) of the FBP-reconstructed CT images are obtained from each downsampling stage. These features are used for the subsequent feature fusions.

Sinogram branch

This component corresponds to the OP_S operator in Eq. [1]. A fully connected layer is employed to simulate the domain transformation to convert the projection domain onto the image domain (27). To work with large receptive fields, dilated convolutions with different dilation rates (d =1, d =6, d =12, and d =18) are applied. The receptive field of a dilated convolution with a kernel size k_d and dilation rate d is equivalent to a convolution with a kernel size of k_c, as follows:

$$k_c = k_d + (k_d - 1)*(d - 1) \quad [2]$$

In addition, the average pooling is also used in parallel to maintain the location structure information of the original input. Next, these 5 features (4 feature maps obtained from convolutions and 1 feature map from

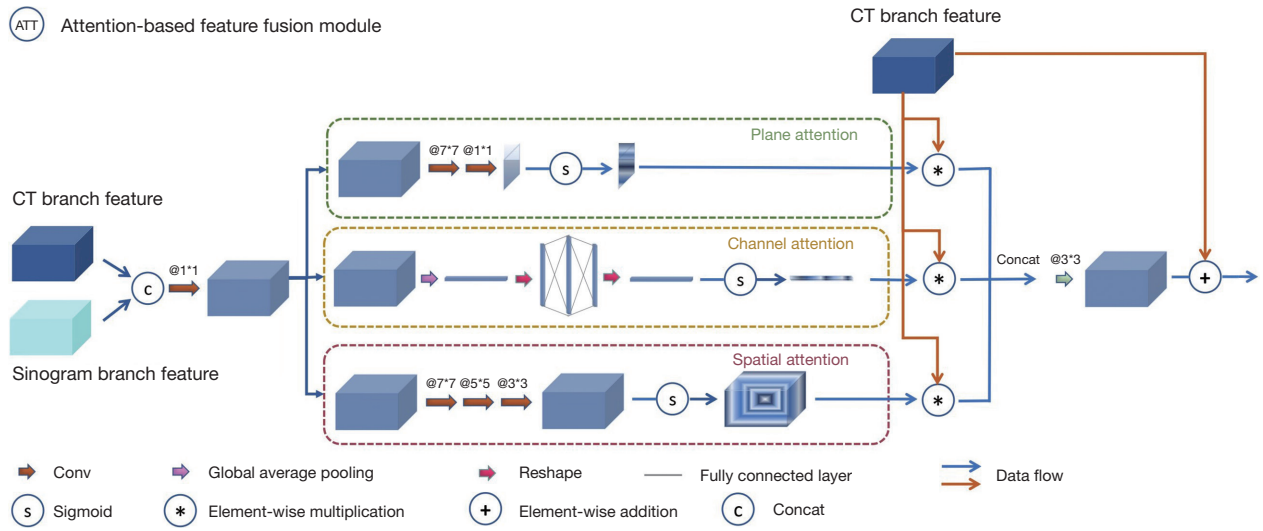

Figure 2 The structure of the attention-based feature fusion module. Three different attention mechanisms are used: plane attention, channel attention, and spatial attention. The symbol @ indicates the convolution kernel size. CT, computed tomography.

average pooling) are cascaded along the depth dimension. The final multiscale sinogram feature with sizes of $(w_1,h_1,c_1),(w_2,h_2,c_2),and\,(w_3,h_3,c_3)$ are sequentially extracted through convolution, pooling, and shortcut connection.

Attention-based feature fusion module

Since 2 types of multiscale feature maps are obtained from 2 feature extractors, the attention-based feature fusion module is used to fuse those features. After this, certain regional information favorable for reconstruction is strengthened, and some irrelevant regional information is weakened.

This component corresponds to the OP_{fusion} operator in Eq. [1] and is denoted as *ATT* in *Figure 1*. Specifically, a convolutional layer with a kernel size of 1×1 is used to compress the fused features from the CT image feature and sinogram feature I_s. The I_f, I_s, and I_{fused} belong to $\mathbb{R}^{W_i\times H_i\times C_i}$, where $W_i,H_i,C_i\in\{w,h,c|(w_1,h_1,c_1),(w_2,h_2,c_2),(w_3,h_3,c_3)\}$. Next, the plane attention module P_{att}, channel attention module D_{att} and spatial attention module S_{att} calculated by I_{fused} are aggregated to generate complete attention maps, which are multiplied with the feature maps extracted from the CT branch. More details are shown in *Figure 2*.

The plane attention module extracts the plane feature map $M_p=\left[n_{1,1},\ldots n_{1,H_i},\ldots n_{W_i,1},\ldots,n_{W_i,H_i}\right]\in\mathbb{R}^{W_i*H_i*1}$ via 2 convolutional layers. The sigmoid function is used to generate the final plane attention map $\hat{M}_p=\left[\hat{n}_{1,1},\ldots\hat{n}_{1,H_i},\ldots\hat{n}_{W_i,1},\ldots,\hat{n}_{W_i,H_i}\right]$.

$$M_p=Conv_{7*7,1*1}\left(I_{fused}\right),\hat{M}_p=\sigma\left(M_p\right) \quad [3]$$

where *Conv* denotes the convolutional layer, and σ denotes the sigmoid activation function. Afterward, the $\hat{M}_p$ is applied to the input FBP feature $I_f=\left[p_{1,1},\ldots p_{1,H_i},\ldots p_{W_i,1},\ldots,p_{W_i,H_i}\right]$ with $p\epsilon\mathbb{R}^{C_i}$ using planewise multiplication to obtain a feature map with different plane weights.

$$I_{f_{PA}}=ATT_p\left(I_f,\hat{M}_p\right)=\left[p_{1,1}\hat{n}_{1,1},\ldots,p_{1,H_i}\hat{n}_{1,H_i},\ldots,p_{W_i,1}\hat{n}_{W_i,1},\ldots,p_{W_i,H_i}\hat{n}_{W_i,H_i}\right] \quad [4]$$

The channel attention module squeezes the feature I_{fused} to obtain $M_c=\left[m_1,m_2,\ldots m_{C_i}\right]\in\mathbb{R}^{C_i}$ via a global average pooling of dimension $\mathbb{R}^{W_i*H_i}$. As shown in the literature, we employ a fully connected layer to extract the global attention features (22,23). The M_c is activated by the sigmoid function to generate the channel attention map $\hat{M}_c=\left[\hat{m}_1,\hat{m}_2,\ldots\hat{m}_{C_i}\right]\in\mathbb{R}^{C_i}$, as follows:

$$M_c=fc\left(A\left(I_{fused}\right)\right),\hat{M}_c=\sigma\left(M_c\right) \quad [5]$$

where fc denotes 2 fully connected layers with total node numbers of $C_i/2$ and C_i, A denotes the global average pooling, and σ denotes the sigmoid activation function. The $\hat{M}_c$ is applied on the input CT feature $I_f=\left[c_1,c_2,\ldots,c_{C_i}\right]$ with $c\epsilon\mathbb{R}^{W_i*H_i}$ using channelwise multiplication to obtain a feature map with different channel weights.

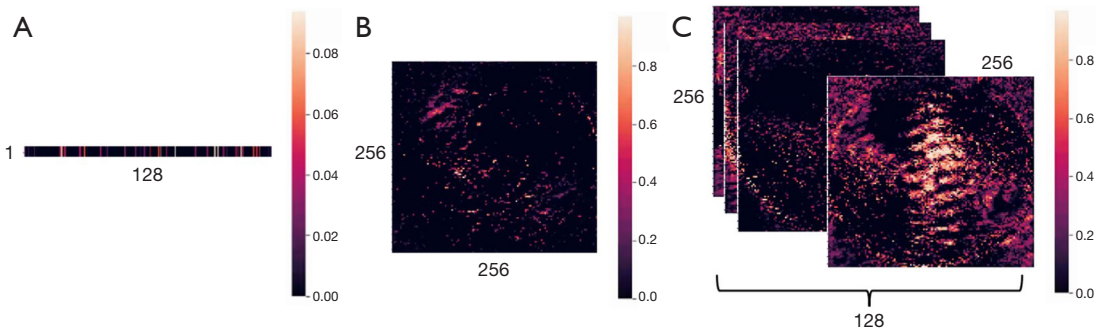

Figure 3 An example of the attention maps ($\in \mathbb{R}^{256*256*128}$) obtained from the proposed ADB-Net network: (A) the channel attention map, (B) the plane attention map, and (C) the spatial attention map.

$$I_{f_{CA}} = ATT_C\left(I_f, \hat{M}_c\right) = \left[c_1\hat{m}_1, c_2\hat{m}_2, ..., c_{C_i}\hat{m}_{C_i}\right] \quad [6]$$

The spatial attention module is used to further enhance the level of attention. The 3-dimensional spatial feature maps $M_s = \left[z_{1,1,1}, \ldots z_{W_i,H_i,C_i}\right] \in \mathbb{R}^{W_i*H_i*C_i}$ are obtained from 3 convolutional layers with a kernel size of 7, 5, and 3. The convolution layer only serves to extract features and does not change the dimensionality. Afterward, the sigmoid activation function is used to generate the $\hat{M}_s = \left[\hat{z}_{1,1,1}, \ldots \hat{z}_{W_i,H_i,C_i}\right] \in \mathbb{R}^{W_i,H_i,C_i}$ map:

$$M_s = Conv_{7*7,5*5,3*3}\left(I_{fused}\right), \hat{M}_s = \sigma\left(M_s\right) \quad [7]$$

The $\hat{M}_s$ is applied to the input CT feature $I_f = \left[s_{1,1,1}, \ldots, s_{W_i,H_i,C_i}\right]$ with $s \epsilon \mathbb{R}^1$ using spatialwise multiplication to obtain a feature map with different spatial weights.

$$I_{f_{SA}} = ATT_S\left(I_f, \widehat{M}_s\right) = \left[s_{1,1,1}\hat{z}_{1,1,1}, \ldots, s_{W_i,H_i,C_i}\hat{m}_{W_i,H_i,C_i}\right] \quad [8]$$

Finally, these unique attention features obtained from the above 3 different attention modules are superimposed along the depth dimension, and convolution is used to fuse the complex information. As depicted in *Figure 3*, the 3 attention maps have unique weight distributions, demonstrating that the attention module plays a critical role in capturing certain features.

Network training

Network loss function

The network loss is a weighted summation of the root-mean-square error (RMSE), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR), as seen in Eq. [9]:

$$L_{loss} = \lambda_1 L_{rmse} + \frac{\lambda_2}{L_{ssim}} + \frac{\lambda_3}{L_{psnr}} \quad [9]$$

where λ_1, λ_2, and λ_3 present the corresponding weights, whose values are empirically determined to balance the 3 loss terms in Eq. [9]. RMSE is used to control the global difference, SSIM is used to control the structural information, and PSNR is used to control the degree of noise. The dynamic range used in the PSNR estimation corresponds to the difference between the maximum and the minimum pixel values of the entire image (32-bit float format).

Network training details

For network training, the Adam optimizer was used, and the learning rate was initialized by 1×10^{-4} with a decay rate of 0.95 after every epoch. The network was trained by 150 epochs on a single Nvidia RTX A4000 GPU card (NVIDIA Corp., Santa Clara, CA, USA). The entire network training took approximately 16 hours.

Dataset

Numerical data

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The data were prepared from low-dose CT images that were published by the

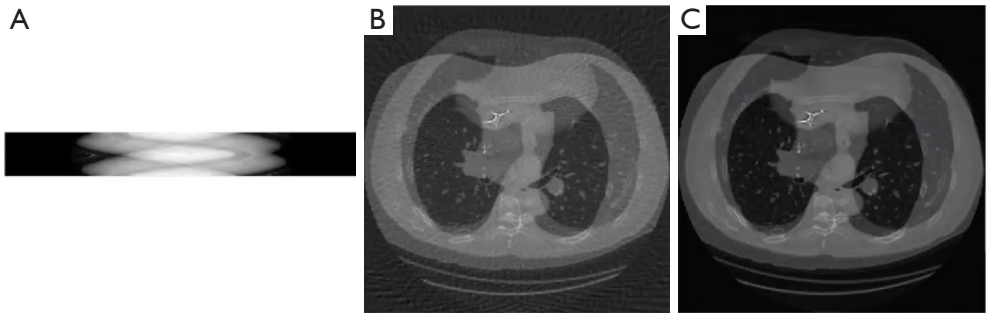

Figure 4 Visualization of a mixup. (A) A mixed sinogram with 128-view of projections. (B) A mixed CT image reconstructed from image A. (C) Ground truth of the mixed CT image in B. CT, computed tomography.

American Association of Physicists in Medicine (AAPM) low-dose CT challenge (28). The forward projections and CT image reconstructions were performed in Python (Python Software Foundation, Wilmington, DE, USA) with self-developed operators. Specifically, the distance from the X-ray source to the rotation center was 1,000 mm, and the distance to the detector was 1,200 mm. There were 1,024 detector elements with an element size of 0.6 mm × 0.6 mm. The pixel size CT image was 0.625 mm × 0.625 mm. In total, 5,410 training dataset and 526 testing dataset were generated. By default, 128 sparse-view projections were simulated.

Experimental data

This study was approved by the Institutional Animal Care and Use Committee (IACUC) of the Shenzhen Institute of Advanced Technology at the Chinese Academy of Sciences and was conducted in compliance with the protocol (SIAT-IACUC-201228-YGS-LXJ-A1498; January 5, 2021) for the care and use of animals. Experimental data were acquired from an in-house CT imaging bench. The system was equipped with a rotating-anode X-ray tube (G-242; Varex Imaging Corp., Salt Lake City, UT, USA) and a flat panel detector (4343CB; Varex Imaging Corp.). The effective detector pixel size was 417 μm × 417 μm. The distance from the X-ray focal spot to the rotation center was 1,156.3 mm, and the distance to the detector plane was 1,560.6 mm. The full-view CT scan had 900 projections with an angular interval of 0.4 degrees. The sparse-view CT scans had 128 projections. An anthropomorphic thorax phantom (Model RS-111 T, Radiology Support Devices, Long Beach, CA, USA) and an *in vivo* anesthetized monkey were scanned.

Data augmentation

A mixup was used to augment the training data, which has been shown to improve the robustness and generalization of the model (29-31). To do so, 2 individual images were mixed randomly over the network training with a certain weight:

$$F_{mixed} = \lambda * F_1 + (1-\lambda) * F_2 \quad [10]$$

$$S_{mixed} = \lambda * S_1 + (1-\lambda) * S_2 \quad [11]$$

$$G_{mixed} = \lambda * G_1 + (1-\lambda) * G_2 \quad [12]$$

where λ denotes the weight coefficient sampled from the beta distribution; and F, S, and G denote the CT image, sinogram, and label, respectively. All training data had a 50% chance of being mixed up with the other training data. One example of image mixup is shown in *Figure 4*. It should be noted that the data mixup was not implemented on the test data. The mixup operation enriched the patterns of the streak artifacts and benefitted the network by removing them more efficiently.

Results

Ablation experiments

Ablation experiments were performed to evaluate the effectiveness of the key components of ADB-Net. The quantitative results are shown in *Table 1*. The baseline is the proposed ADB-Net without any modification. In this case, RMSE is 20.6160, SSIM is 0.9257, and PSNR is 38.8246.

Dual branch

As shown in *Figures 1*,*2* different network branches were designed to extract the CT image features and the sinogram features from the sinogram input at the same time. These 2 different domain features were fused to

Table 1 Quantitative comparison results of the ablation experiments

Network ablated modules	Performance		
	RMSE	SSIM	PSNR
ADB-Net (ablation)			
Dual-branch	21.9263±0.1766	0.9200±0.0022	38.2921±0.1904
Fully connected layer	20.9046±0.1815	0.9244±0.0022	38.7024±0.1934
Dilated convolution	21.0535±0.1728	0.9242±0.0021	38.6362±0.1920
Attention fusion	21.4903±0.1757	0.9231±0.0022	38.4575±0.1915
Channel attention	20.8785±0.1825	0.9247±0.0023	38.7130±0.1942
Spatial attention	20.9421±0.1818	0.9242±0.0022	38.6856±0.1955
Plane attention	21.3234±0.1775	0.9231±0.0022	38.5282±0.1939
Mixup	21.3506±0.1755	0.9227±0.0022	38.5144±0.1913
ADB-Net	20.6160±0.1822	0.9257±0.0021	38.8246±0.1943

The values of RMSE, PSNR, and SSIM are estimated from the 526 test samples. The values are presented as mean ± standard deviation. ADB-Net, attention-based dual-branch network; RMSE, root-mean-square error; SSIM, structural similarity; PSNR, peak signal-to-noise ratio.

calculate the attention map at the attention fused module. To demonstrate the necessity of the 2 domain inputs of the attention module, the sinogram feature were replaced by the CT image features. Specifically, the features of CT were copied, and the 2 identical CT features were put into the attention-based feature fusion module. According to the results in *Table 1*, it is easy to see that the CT image features and the sinogram features were all needed to improve the network performance.

Fully connected layer

The fully connected layer in the sinogram feature extraction module was removed to verify its importance. All the other structures were unchanged. As shown in *Table 1*, the performance would degrade without the fully connected layer. One possible explanation for this was that the sinogram and FBP belonged to different image domains resulting in different representations of the same information, so directly fused features might introduce some small errors. The global operation, for example, a fully connected layer, can help to eliminate the difference between projection domain feature space and image domain feature space.

Dilated convolution

In this part, the dilated convolution and the pooling operation (ASPP module) in the ADB-Net were removed. As shown in *Table 1*, the ADB-Net without the dilated convolution showed worse performance than original complete ADB-Net. One possible reason for this was that the streaking artifacts were nonlocal and needed to be mitigated via global operations. Dilated convolution has a larger receptive field than traditional convolution and can better handle such problems.

Attention fusion

The attention fusion module played the role of enhancing the extractions of streaking artifact feature information. As clearly shown by the results listed in the fourth and sixth rows of *Table 1*, the use of the attention mechanism can boost the performance of ADB-Net in helping the intermediate layers be more focused on extracting the artifact-related features.

Components of the attention module

The indispensability of each component (plane attention, spatial attention, and channel attention) of the attention module was verified. As shown in the fifth to the seventh row in *Table 1*, any structure removal caused a loss of precision.

Mixup

Compared to the conventional data augmentation approaches, such as image scaling and image rotation,

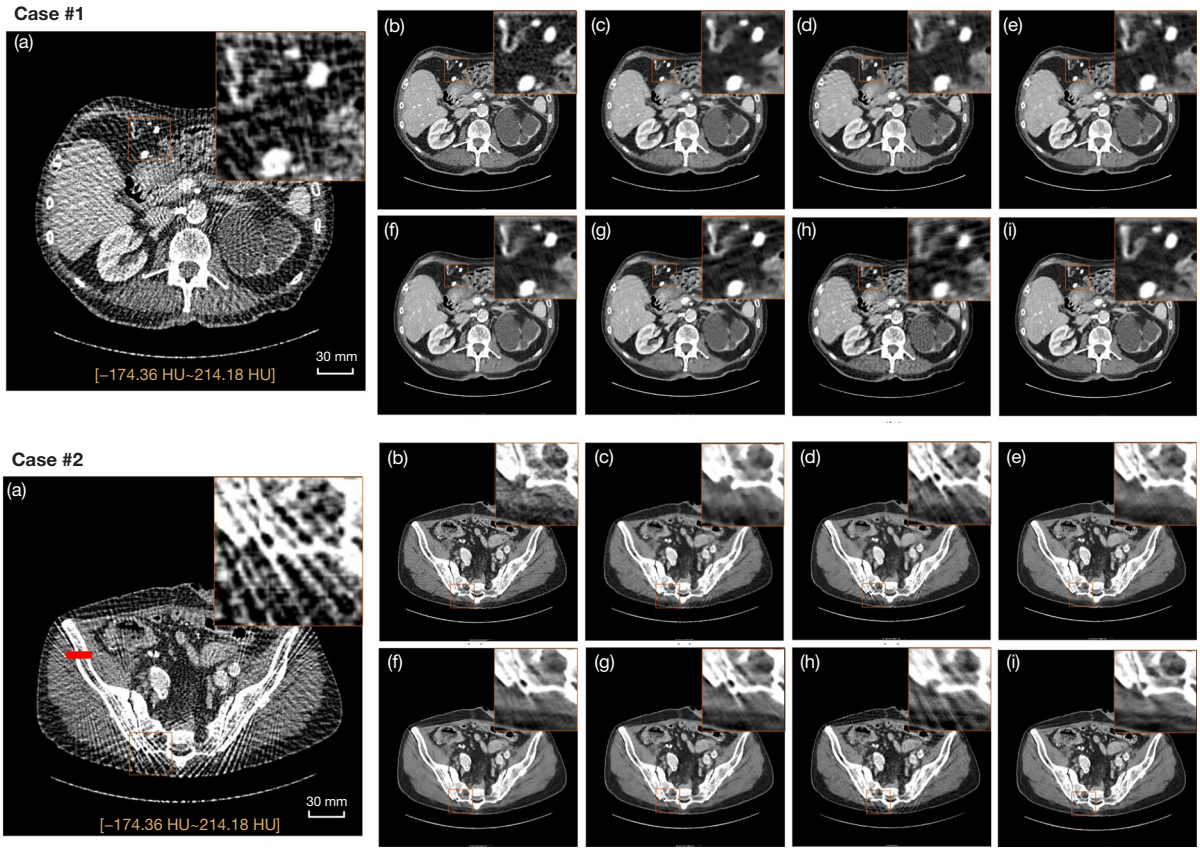

Figure 5 Results of simulated CT images for different reconstruction methods: (a) FBP, (b) ground truth, (c) TV, (d) Red-CNN, (e) FBPConvNet, (f) Frame U-Net, (g) DDNet, (h) SS-Net, and (i) ADB-Net. The display window is [–173.36, 214.18] HU. The scale bar denotes 30 mm. The red line in Case #2 (a) highlights the high-contrast pixels for profile measurements in *Figure 6*. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network; HU, Hounsfield unit.

the image mixup can improve the model's capability in capturing the artifacts distributions and enhance its robustness to different object structures. As shown in *Table 1*, the training data mixup resulted in an improvement of 0.7346, 0.003 and 0.3102 in RMSE, SSIM, and PSNR, respectively.

Comparison experiments

The total variation (TV) (5), FBP convolutional network (FBPConvNet) (8), residual encoder-decoder CNN (Red-CNN) (32), DDNet (10), Frame U-Net (9), SS-Net (16) were used as the comparison methods to the ADB-Net. Essentially, these algorithms can be divided into 3 categories: the TV method belongs to the iterative algorithm; the FBPConvNet, Red-CNN, DDNet, and Frame U-Net methods represent the image domain-only postprocessing strategy; and the SS-Net method represents the sinogram domain-only postprocessing strategy.

Numerical results

The numerical performance of ADB-Net was validated on the AAPM low-dose CT data. The results are presented in *Figure 5*. The TV algorithm removed most of the streaking artifacts but blurred the fine structures. The methods based on either the CT image or the sinogram had difficulty removing substantial streaks while preserving the fine details of the image. However, the streaking artifacts could be mitigated by the proposed ADB-Net without sacrificing the image's sharpness. In addition, line profiles of the high-

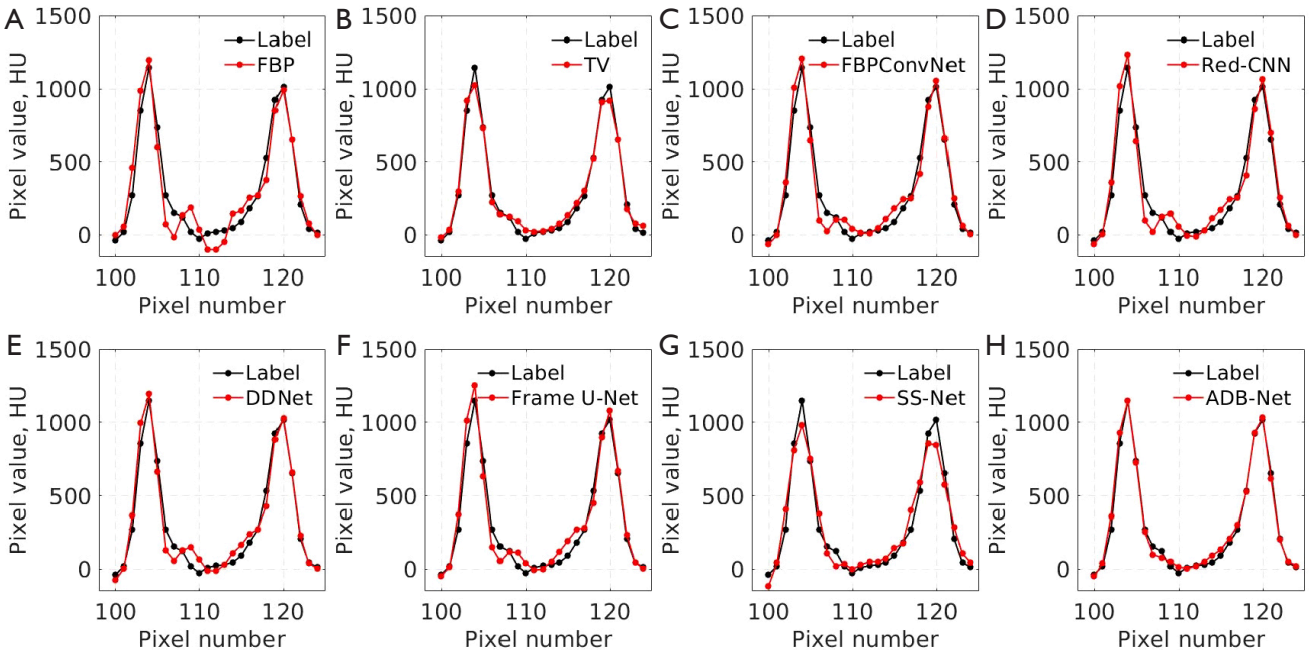

Figure 6 The profiles along the horizontal red line in Figure 5 [image (a) of case #2]. The black curves denote the ground truth, and the red curves denote the different methods: (A) FBP, (B) TV, (C) FBPConvNet, (D) Red-CNN, (E) DDNet, (F) Frame U-Net, (G) SS-Net, and (H) ADB-Net. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network.

contrast bone (highlighted in the red line in *Figure 5* image (a) of case #2) were also compared (*Figure 6*). The ADB-Net method generated the most consistent profile compared to the ground truth, indicating that it had the best capability of maintaining the signal precision and image sharpness of all the methods. The quantitative analysis results are listed in *Table 2*. Again, the ADB-Net achieved the smallest RMSE value and the highest SSIM and PSNR values. The use of a fully connected layer led to a considerable increase in the numbers of parameters. As a consequence, the ADB-Net had a parameter size of 275.88 M. In contrast, the FBPConvNet had a parameter size of 22.08 M, Red-CNN had a parameter size of 18.49 M, DDNet had a parameter size of 25.09 M, Frame U-Net had a parameter size of 25.09 M, and SS-Net had a parameter size of 34.51 M.

Experimental results

Two sets of experimental sparse-view CT imaging data, one each from an anthropomorphic thorax phantom and an anesthetized monkey, were verified (*Figures 7,8*). With regard to the selected region of interest in *Figure 7*, the proposed ADB-Net outperformed the other methods in removing the streaking artifacts. Although the TV and Frame U-Net methods could reduce more artifacts, the reconstructed images became quite blurry. For the animal experiments, again, the ADB-Net showed the best capability in eliminating the streaking artifacts, as shown in the regions highlighted by the white arrows in *Figure 8*.

Moreover, the image spatial resolution—that is, the modulation transfer function (MTF)—was compared quantitatively (*Figure 9*). MTF is an evaluation method of image sharpness, which is related to the contrast and resolution of the image (33-35). The MTF curves were calculated from the highlighted bone or tissue region as shown by the red line in *Figure 7*. To do so, the corresponding edge-spreading profiles were first measured. Second, the line spreading profiles were obtained after differentiation. Finally, the MTF curve was generated from the Fourier transformation. Overall, the MTF curves further confirmed the visual performance shown in *Figures 7,8*. For example, the MTF curve generated from the TV algorithm became narrower, indicating a certain loss of image resolution. Compared with the reference FBP method (for full-view reconstruction), the ADB-Net

Table 2 Quantitative comparison results for the 128-view CT imaging

Method	Group mean			Case #1			Case #2		
	RMSE	SSIM	PSNR	RMSE	SSIM	PSNR	RMSE	SSIM	PSNR
FBP	74.6665	0.5466	27.6217	55.4239	0.5350	27.4446	61.9825	0.4706	25.0570
TV (5)	22.5669	0.9238	38.0336	17.1821	0.8971	36.5167	15.7972	0.9064	35.5981
FBPConvNet (8)	23.0779	0.9120	37.8346	17.2667	0.8783	36.0945	14.7986	0.8996	35.6193
Red-CNN (32)	24.8279	0.9019	37.1899	17.9570	0.8672	35.5857	15.5983	0.8859	34.9057
DDNet (10)	22.8952	0.9143	37.9064	17.3675	0.8802	36.1204	14.8013	0.9027	35.6625
Frame U-Net (9)	22.7876	0.9144	37.9458	17.4995	0.8806	36.1178	15.1155	0.9014	35.6786
SS-Net (16)	37.3478	0.8497	33.6550	25.9916	0.8110	32.3871	23.1603	0.8302	31.9081
ADB-Net	20.6160	0.9257	38.8246	15.9599	0.8924	36.7648	13.1468	0.9159	36.7692

The averaged performance of the 526 testing images and the performance of the 2 selected cases in *Figure 5* were evaluated. The second to fourth columns are the average of the total test set, while the fifth to seventh and eighth to tenth columns are the results of 2 cases in the test set. RMSE, root-mean-square error; SSIM, structural similarity; PSNR, peak signal-to-noise ratio. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network; CT, computed tomography.

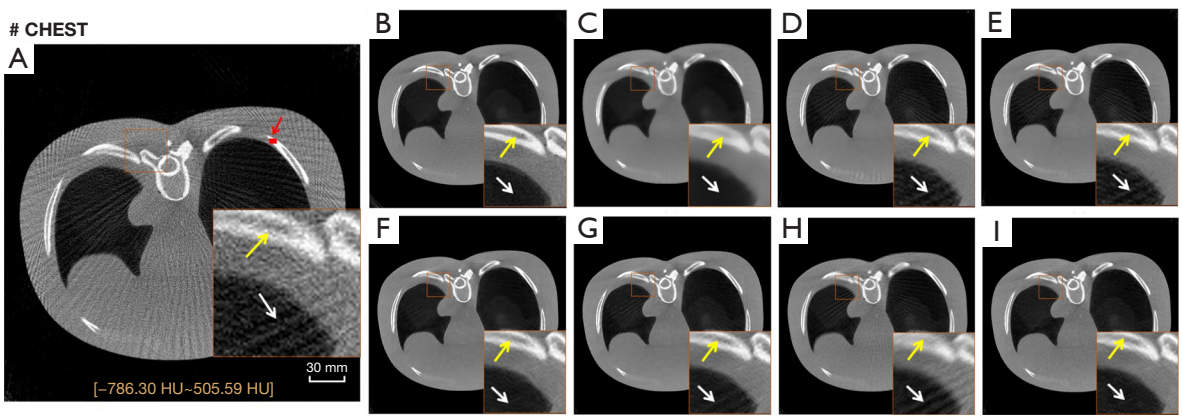

Figure 7 Experimental results of the thorax phantom: (A) FBP, (B) ground truth, (C) TV, (D) Red-CNN, (E) FBPConvNet, (F) Frame U-Net, (G) DDNet, (H) SS-Net, and (I) ADB-Net. The display window is [–786.30, 505.59] HU. The scale bar denotes 30 mm. The yellow arrows highlight the preservation of structure by each method, while the white arrows highlight the ability to remove streaking artifacts. The red line pointed at by the red arrow indicates the region used in the MTF calculation. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network; MTF, modulation transfer function.

method slightly degraded the image resolution. Note that the MTF was measured on the high-contrast object; thus, the actual visual performance might vary for low-contrast objects.

Results of the robustness study

The robustness of the proposed ADB-Net in removing the streaking artifacts over different sparse-view projection data (i.e., 64, 256, 384, and 512) was investigated (*Figure 10*). Since the fully connected layer required a fixed input size, we could not directly use the model trained on 128 views

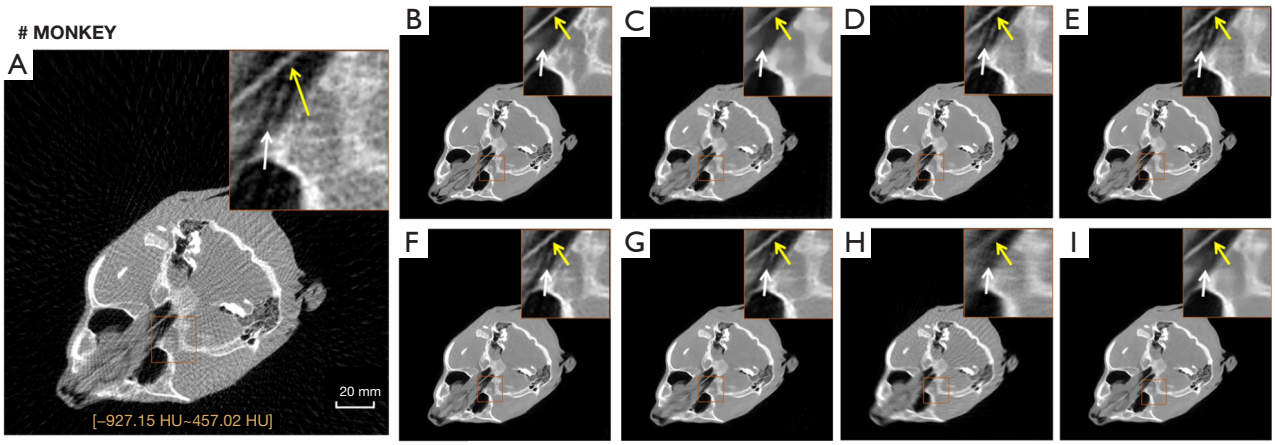

Figure 8 Experimental results of the monkey head: (A) FBP, (B) ground truth, (C) TV, (D) Red-CNN, (E) FBPConvNet, (F) Frame U-Net, (G) DDNet, (H) SS-Net, and (I) ADB-Net. The display window is [–927.15, 457.02] HU. The scale bar denotes 20 mm. The yellow arrows highlight the preservation of structure of each method, while the white arrows highlight the ability to remove streaking artifacts. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep-neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network.

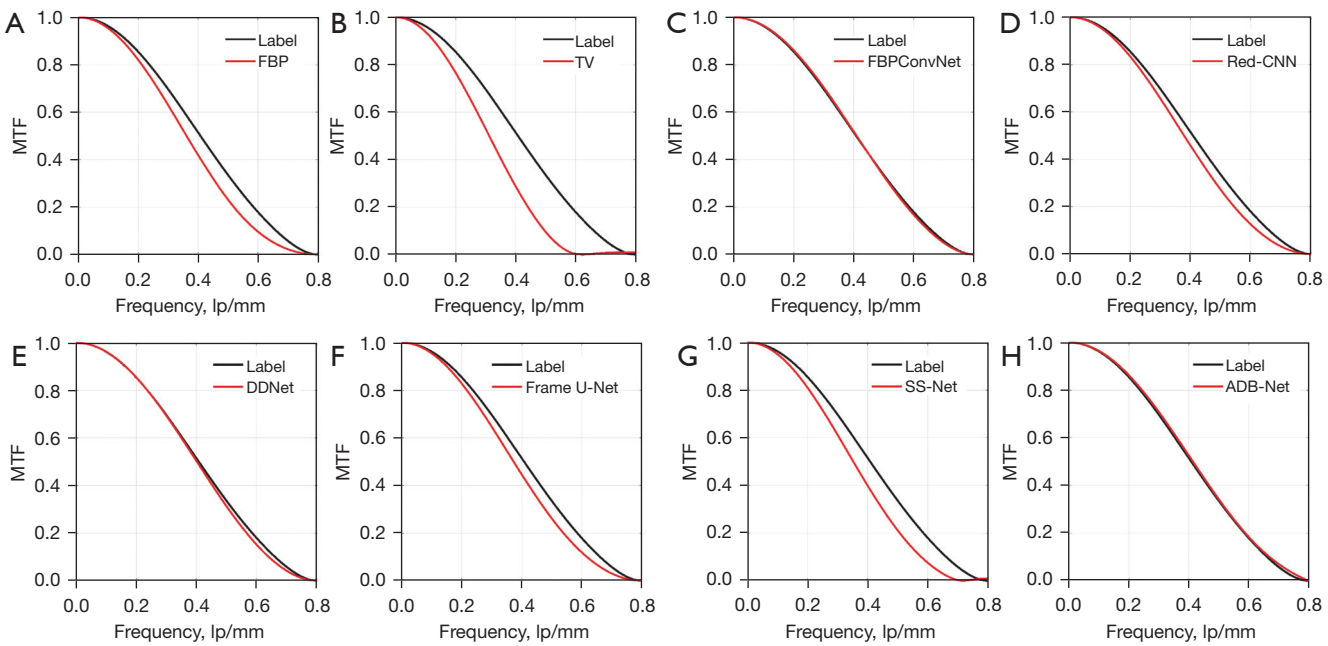

Figure 9 The MTF results. (A) FBP, (B) TV, (C) FBPConvNet, (D) Red-CNN, (E) DDNet, (F) Frame U-Net, (G) SS-Net, and (H) ADB-Net. The results of ADB-Net and FBPConvNet are approximate to the label curve, while the results of the other methods are not very satisfactory. Note that the MTF is measured on the high-contrast object; thus, the actual visual performance may vary for low-contrast objects. FBP, filtered back-projection; TV, total variation; FBPConvNet, FBP convolutional network; Red-CNN, residual encoder-decoder convolutional neural network; DDNet, DenseNet and deconvolution-based network; Frame U-Net, dual-frame U-Net via deep convolutional framelets; SS-Net, deep neural network-enabled sinogram synthesis method; ADB-Net, attention-based dual-branch network; MTF, modulation transfer function.

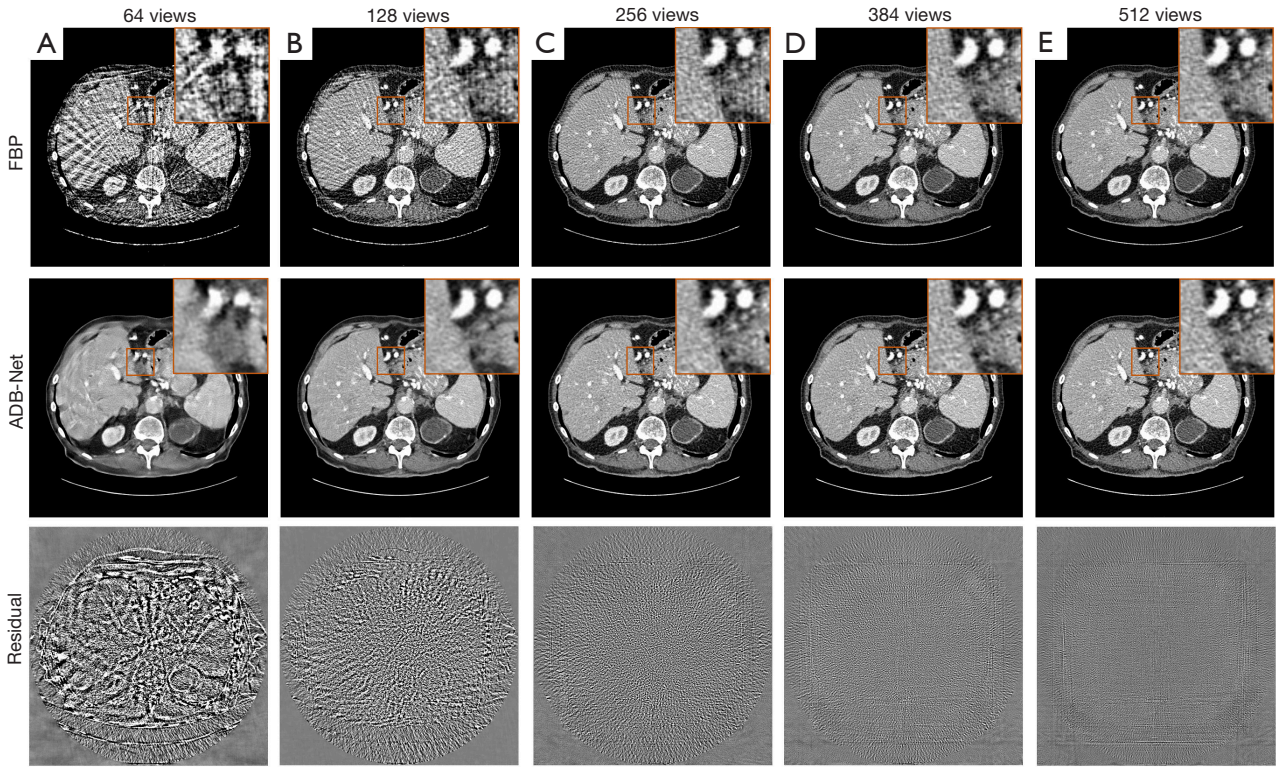

Figure 10 The results of different numbers of projections. Images in the first row are reconstructed by the FBP algorithm, and images in the second row are reconstructed by the ADB-Net. The residual images of the ground truth and ADB-Net are depicted in the third row. The display window is [−173.36, 214.18] HU for the CT images and [−50.00, 50.00] HU for the residual images. (A) 64 views, (B) 128 views, (C) 256 views, (D) 384 views, and (E) 512 views. ADB-Net, attention-based dual-branch network; FBP, filtered back-projection; HU, Hounsfield unit; CT, computed tomography.

to test the other data with other views. As a result, extra convolutional layers were added to make the features of different views consistent before feeding the data into the fully connected layer. During the network training, only the added convolutional layer was trained, while the other parameters (trained by 128 views) were fixed. As observed, the image quality became less satisfactory when the number of projections was 64. However, ADB-Net was still able to remove most of the streaking artifacts and keep the majority of structural information intact under this very challenging condition. As the number of projections increased, the quality of the reconstructed CT images from ADB-Net gradually improved, and the residual losses were found to be negligible. The quantitative analysis results are listed in *Table 3*. It should be noted that certain convolutions and pooling layers were altered in the sinogram feature extraction module for those sinograms with different dimensions to keep consistent feature sizes.

Discussion

In this study, a novel attention-based dual-branch end-to-end network, ADB-Net, was proposed to reconstruct high-quality CT images directly from the sparse-view sinogram. This model extracts the information stored in the CT image and the sinogram using 2 parallel branches and fuses them through attention mechanisms. By doing so, complement information is accessed to assist the sparse-view CT image reconstruction. To generate the best imaging performance, certain techniques are used to enhance the feature extractions. For instance, the fully connected layer is applied to reduce the difference in feature space between the projection domain and the image domain, and the dilated convolutional layer is used to enlarge the receptive field. The superior performance of the ADB-Net network was evaluated via additional ablation experiments. Finally,

Table 3 Quantitative comparison results for different views [64, 128, 256, 384, 512]

Views	FBP/ADB-Net		
	RMSE	SSIM	PSNR
64	135.0468/32.7989	0.3375/0.8641	22.4709/34.7748
128	74.6665/20.6160	0.5466/0.9257	27.6217/38.8246
256	38.4785/15.0697	0.7838/0.9578	33.3739/41.5471
384	22.7170/11.4356	0.8973/0.9748	37.9521/43.9424
512	17.7779/9.1669	0.9389/0.9832	40.0803/45.8646

The values on the left represent the quantitative results of FBP, and the values on the right represent the results of the ADB-Net prediction. FBP, filtered back-projection; ADB-Net, attention-based dual-branch network; RMSE, root-mean-square error; SSIM, structural similarity; PSNR, peak signal-to-noise ratio.

numerical simulations, phantoms, and animal experiments were conducted. The results demonstrated that this newly developed network could consistently remove the streaking artifacts while maintaining the fine structures.

The ADB-Net trained with numerical data was directly used on the experimental data without any network fine-tuning. Due to this reason, the reconstructed experimental CT images might show inferior performance to the numerical simulations. Enhanced performance of the ADB-Net network would be expected if further network fine-tuning could be implemented on real experimental data.

In the future, a number of topics can be investigated. First, the self-attention mechanism defined in the transformer network could be used to carry out the global feature extractions, particularly for the sinogram domain subnetwork (36,37). By doing so, removing the streaking artifacts that spread over the entire CT image could become more effective. Second, the hint learning approach could be tested and incorporated to replace the fully connected layer in the sinogram domain subnetwork with the purpose of greatly shrinking the total size of the network parameters while minimally degrading the entire network performance (38). Third, the capability of performing high-quality CT image reconstruction from a truncated sinogram (super-short scan) could be investigated with a modified ADB-Net (39).

Conclusions

We propose an attention-based dual-branch end-to-end deep network, ADB-Net, to reconstruct high-quality sparse-view CT images. The performance of ADB-Net was validated on numerical simulations, an anthropomorphic thorax phantom, and *in vivo* preclinical experiments. Results demonstrate that this newly developed network can remove the streaking artifacts robustly while maintaining the fine structural details in sparse-view low-dose CT imaging.

Acknowledgments

The authors would like to thank Dr. Yaoqin Xie for lending us the anthropomorphic thorax phantom.
Funding: This work was supported by the National Natural Science Foundation of China (No. 62201560, 12027812), the Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515110685 and 2021A1515111031), the Shenzhen Basic Research Program (No. JCYJ20200109115212546), the Guangdong Key Area Research Program (No. 2020B1111130001), and the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2021362).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-609/coif). DL serves as an unpaid editorial board member of *Quantitative Imaging in Medicine and Surgery*. The other 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). The animal experiment was approved by the Institutional Animal Care and Use Committee (IACUC) of the Shenzhen Institute of Advanced Technology at the Chinese Academy of Sciences and was conducted in compliance with the protocol (SIAT-IACUC-201228-YGS-LXJ-A1498; January 5, 2021) for the care and use of animals.

References

1. 1990 Recommendations of the International Commission on Radiological Protection. Ann ICRP 1991;21:1-201.
2. Gu J, Ye JC. AdaIN-based tunable CycleGAN for efficient unsupervised low-dose CT denoising. IEEE Transactions on Computational Imaging 2021;7:73-85.
3. Liang K, Zhang L, Yang Y, Xing Y. A self-supervised deep learning network for low-dose CT reconstruction. 2018 IEEE nuclear science symposium and medical imaging conference proceedings (NSS/MIC). Sydney: IEEE 2018:1-4.
4. Ding Q, Nan Y, Gao H, Ji H. Deep learning with adaptive hyper-parameters for low-dose CT image reconstruction. IEEE Transactions on Computational Imaging 2021;7:648-60.
5. Wu W, Hu D, An K, Wang S, Luo F. A high-quality photon-counting CT technique based on weight adaptive total-variation and image-spectral tensor factorization for small animals imaging. IEEE Transactions on Instrumentation and Measurement 2020;70:1-14.
6. Xu Q, Yu H, Mou X, Zhang L, Hsieh J, Wang G. Low-dose X-ray CT reconstruction via dictionary learning. IEEE Trans Med Imaging 2012;31:1682-97.
7. Zhang H, Huang J, Ma J, Bian Z, Feng Q, Lu H, Liang Z, Chen W. Iterative reconstruction for x-ray computed tomography using prior-image induced nonlocal regularization. IEEE Trans Biomed Eng 2014;61:2367-78.
8. Kyong Hwan Jin, McCann MT, Froustey E, Unser M. Deep Convolutional Neural Network for Inverse Problems in Imaging. IEEE Trans Image Process 2017;26:4509-22.
9. Han Y, Ye JC. Framing U-Net via Deep Convolutional Framelets: Application to Sparse-View CT. IEEE Trans Med Imaging 2018;37:1418-29.
10. Zhang Z, Liang X, Dong X, Xie Y, Cao G. A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution. IEEE Trans Med Imaging 2018;37:1407-17.
11. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE 2017.
12. Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med Phys 2017;44:e360-75.
13. Kang E, Min J, Ye JC. Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction. arXiv 2017. arXiv:1703.01383.
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016:770-8.
15. 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.
16. Lee H, Lee J, Kim H, Cho B, Cho S. Deep-neural-network-based sinogram synthesis for sparse-view CT image reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences 2018;3:109-19.
17. Li Y, Li K, Zhang C, Montoya J, Chen GH. Learning to Reconstruct Computed Tomography Images Directly From Sinogram Data Under A Variety of Data Acquisition Conditions. IEEE Trans Med Imaging 2019;38:2469-81.
18. Fu L, De Man B. A hierarchical approach to deep learning and its application to tomographic reconstruction. 15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine. Proc. SPIE 2019;11072:1107202.
19. Lin WA, Liao H, Peng C, Sun X, Zhang J, Luo J, Chellappa R, Zhou SK. DuDoNet: Dual Domain Network for CT Metal Artifact Reduction. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, 2019.
20. Wang C, Shang K, Zhang H, Li Q, Hui Y, Zhou SK. DuDoTrans: Dual-Domain Transformer Provides More Attention for Sinogram Restoration in Sparse-View CT Reconstruction. arXiv 2021. arXiv:2111.10790.
21. Sun C, Deng K, Liu Y, Yang H. A Lightweight Dual-Domain Attention Framework for Sparse-View CT Reconstruction. arXiv 2022. arXiv:2202.09609.
22. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. editors. ECCV 2018. Lecture Notes in Computer Science. Cham: Springer, 2018.
23. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H. Dual Attention Network for Scene Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, 2019.
24. Hu R, Yan H, Nian F, Mao R, Li T. Unsupervised computed tomography and cone-beam computed tomography image registration using a dual attention

network. Quant Imaging Med Surg 2022;12:3705-16.
25. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. editors. ECCV 2018. Lecture Notes in Computer Science. Cham: Springer, 2018.
26. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. editors. MICCAI 2015. Lecture Notes in Computer Science. Cham: Springer, 2015.
27. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature 2018;555:487-92.
28. AAPM low dose CT grand challenge. Available online: https://www.aapm.org/grandchallenge/lowdosect/
29. Zhang H, Cisse M, Dauphin YN, David Lopez-Paz D. Mixup: Beyond empirical risk minimization. arXiv 2017. arXiv:1710.09412.
30. Huang Z, Tang S, Chen Z, Wang G, Shen H, Zhou Y, Wang H, Fan W, Liang D, Hu Y, Hu Z. TG-Net: Combining transformer and GAN for nasopharyngeal carcinoma tumor segmentation based on total-body uEXPLORER PET/CT scanner. Comput Biol Med 2022;148:105869.
31. Zhang L, Deng Z, Kawaguchi K, Ghorbani A, Zou J. How does mixup help with robustness and generalization? arXiv 2020. arXiv:2010.04819.
32. 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.
33. Hufnagel RE, Stanley NR. Modulation Transfer Function Associated with Image Transmission through Turbulent Media. J Opt Soc Am A 1964;54:52-61.
34. Rossmann K. Point spread-function, line spread-function, and modulation transfer function. Tools for the study of imaging systems. Radiology 1969;93:257-72.
35. Ge Y, Zhang Q, Hu Z, Chen J, Shi W, Zheng H, Liang D. Deconvolution-Based Backproject-Filter (BPF) Computed Tomography Image Reconstruction Method Using Deep Learning Technique. arXiv 2018. arXiv:1807.01833.
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: NIPS'17, 2017.
37. Liang J, Yang C, Zeng M, Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. Quant Imaging Med Surg 2022;12:2397-415.
38. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for Thin Deep Nets. arXiv 2014. arXiv:1412.6550.
39. Podgorsak AR, Shiraz Bhurwani MM, Ionita CN. CT artifact correction for sparse and truncated projection data using generative adversarial networks. Med Phys 2021;48:615-26.

Cite this article as: Gao X, Su T, Zhang Y, Zhu J, Tan Y, Cui H, Long X, Zheng H, Liang D, Ge Y. Attention-based dual-branch deep network for sparse-view computed tomography image reconstruction. Quant Imaging Med Surg 2023;13(3):1360-1374. doi: 10.21037/qims-22-609

