Low Dose CT Image Denoising Using Efficient Transformer with SimpleGate Mechanism

Lianjin Xiong[1], Wei Qiu[1], Ning Li[1], Yishi Li[1,2], and Yangsong Zhang[1](✉)

[1] School of Computer Science and Technology, Laboratory for Brain Science and Medical Artificial Intelligence, Southwest University of Science and Technology, Mianyang 621010, China
zhangysacademy@gmail.com

[2] Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Abstract. The widespread use of computed tomograph (CT) technology in clinic has caused more and more patients to worry that they will receive too much radiation during the scanning. The low-dose CT (LDCT) scanning is more likely to be accepted by the patients. But LDCT images can adversely affect doctors' diagnosis, owing to low quality of the images. Therefore, it is necessary to improve the diagnostic performance by denoising LDCT images. During the past few decades, the convolutional neural networks (CNNs) and Transformer models that achieve remarkable performance in natural image denoising provide new avenues for LDCT denoising. Although the existing methods have successfully achieved noise reduction, there is still large room for improvement in the denoising level. In this paper, we refer to the implementation of natural images denoising, and proposed a transformer-based U-shape network model to achieve denoising in LDCT images. In each transformer block, we used the depth-wise convolution, transposed self-attention mechanism, and SimpleGate to improve performance and speed up efficiency. Extensive experiments on the AAPM-Mayo clinic LDCT Grand Challenge dataset indicated that the proposed model yielded a competitive performance to the compared baseline denoising methods. In particular, good evaluation was achieved in noise suppression, structure preservation and lesion highlighting.

Keywords: Low dose CT · deep learning · denoising · transformer

1 Introduction

Computed tomography (CT) system, as noninvasive imaging equipment, has been widely used for medical diagnosis and treatment [16]. It works by collecting X-rays that pass through the body and then reconstructing each slice of organ tissues. But the radiation from X-rays causes some damage to cells and DNA in the body [1]. We can roughly assume that dose is positively related to risk and imaging quality, high dose means high risk and high imaging quality.

M. Tanveer et al. (Eds.): ICONIP 2022, LNCS 13625, pp. 556–566, 2023.
https://doi.org/10.1007/978-3-031-30111-7_47

The weaker the X-ray flux, the noisier a reconstructed CT image, which could influence the doctor's diagnosis. Therefore, many algorithms have been designed to alleviate this problem. In general, these algorithms can be divided into three categories, i.e., sinogram domain filtration, iterative reconstruction, and image post-processing.

Sinogram filtering techniques process the raw data before image reconstruction. The most famous and widely used algorithm is filtered backprojection (FBP) [18]. However, these methods often suffer spatial resolution loss or artifacts in the reconstructed images. Besides, it is very difficult for ordinary users to get raw data of commercial scanners. Over the past decade, researchers have worked to develop new iterative reconstruction (IR) algorithms in the field of LDCT. In general, these algorithms optimize an objective function, which includes a statistical noise model [19,25], priors information in the image domain and an image system model [8,12]. Although IR techniques obtained excellent results, it is impossible to apply it to practical applications because of the high computational cost. On the other hand, image post-processing is computationally efficient and directly perform on an image. Many methods have played a certain role in suppressing noise and artifacts, such as non-local means (NLM) method [15], dictionary-learning-based K-SVD method [7] and block-matching 3D (BM3D) algorithm [10,11].

In the past few years, deep learning technologies extensively applied in other fields have also attracted tremendous attention in the field of medical images, such as image registration [3,27], image segmentation [30], image classification [4] and LDCT denoising [5,13,21,23,26,28,32]. For example, Chen et al. [5] designed a residual encoder-decoder network for LDCT images, which greatly suppresses the image noise through the method based on CNN. Yang et al. [28] used WGAN network and perceptual loss to promise the reconstructed image quality. They greatly improved the smoothing problem by using perceptual loss, but obtained not very high values of the evaluation metrics PSNR and SSIM. Zhang et al. [31] designed CLEAR based on GAN network to achieve noise reduction very well. They used multi-level consistency loss and got a big improvement in the evaluation of PSNR and SSIM, but added multiple hyperparameters which brought difficulties to training and application. Liang et al. [13] used the soble operator to extract the edge information of the image, and then used convolution and dense connections to achieve denoising. And recently, Transformer has made great progress in the natural language processing (NLP) fields [22]. This also brings new research ideas to the computer vision (CV) fields. In particular, Vision Transformer (ViT) designed by Dosovitskiy et al. almost makes Transformer the main method in CV field [9]. Many computer vision tasks have reached the state-of-the-art performance through the use of Transformer architecture. However, the Transformer model has not been well applied in the field of LDCT denoising. For example, Zhang et al. proposed a dual-path TransCT network to predict the high quality images by fusing high-frequency and low-frequency features [32]. They successfully brought Transformer to the task of CT denoising, but their model did not outperform CNN by much in the evaluation metric PSNR. Wang et al. [24] proposed a convolution-free T2T vision Transformer to achieve a better noise reduction effect. These Transformer-based methods often require a lot of computational consumption when calculating self-attention. To alleviate the inherent problem and further improve denoising performance, we proposed a new network architecture, which was inspired by the facts that: (1) different methods for image denoising have complementary image prior models ability and can be incorporated to boost the performance [2]; (2) Restormer [29] and NAFNet [6] designed different network structures and achieved very good denoising performance. In the proposed model, we used a transformer block to combine depthwise convolution, transposed self-attention mechanism, and SimpleGate, and then plug it as the main building block into the UNet architecture. We termed the proposed model as SAGformer hereinafter. Extensive experiments on Mayo LDCT dataset demonstrate the superiority of our method over other baseline methods.

2 Method

In this study, our goal is to realize denoising of LDCT images by designing an efficient transformer block. In order to reduce the computational cost, we used the modification scheme of self-attention designed by Zamir et al. [29], and adopted the SimpleGate designed by Chen et al. [6] to replace the activation function. First, we introduced the overall structure of the proposed SAGformer (see Fig. 1). Then we presented the main components of the proposed transformer block: (a) Multi-Dconv Head Transposed Self-Attention (MDTSA) and (b) Dconv SimpleGate Feed-forward network (DSGFN).

Overall Structure. From the perspective of macrostructure, we built a standard UNet [20] structure network with four-level symmetric encoder-decoder. First, we applied a 3×3 convolution to learn low-dimensional features. In the encoder stage, downsampling was added to each layer to reduce the resolution and widen the channels of the image. With the deepening of the number of layers, the number of transformer blocks in each layer gradually increases. The resolution of the last level of encoder image is reduced by one-eighth, and the number of channels is eight times the original. On the other hand, in the decoder phase, upsampling between layers is used to restore information. In addition, we also used skip connections to assist the restoration of information and alleviate the problem of gradient disappearance during training. After concatenating the features of the encoder and the decoder, a 1×1 convolution is added to keep the number of channels unchanged, except the top one. Finally, the reconstructed features are added to the LDCT image to output normal dose CT (NDCT) image. Next, we describe the details of the Transformer block.

2.1 Multi-Dconv Head Transposed Self-attention

In the transformer network, one of the most important components is self-attention, which brings good results for vision tasks while incurring expensive

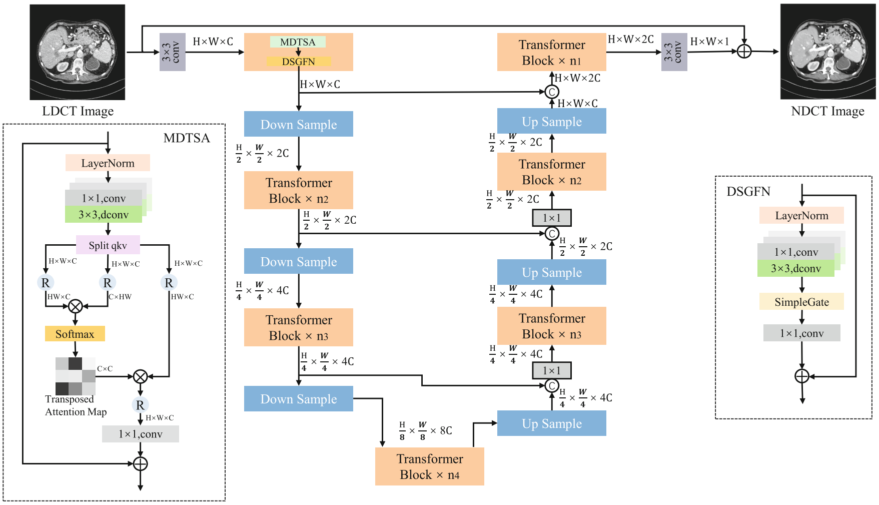


Fig. 1. Architecture of SAGformer for LDCT denoising. The macrostructure is a U-shaped network architecture, in which the transformer module is composed of MDTSA and DSGFN, which mainly realize denoising by performing attention calculation on the channel dimension. R in the figure stands for Reshape operation.

computational costs. The computing mechanism of self-attention (SA) determines that it has a large receptive field and can process global features, but this also leads to its insufficiency in local feature processing [22]. Therefore, we decided to introduce depthwise convolution to enrich local feature information. The reason for using depth-wise convolution is that it can reduce the amount of computation compared to common convolution. But before that, we added a 1 × 1 convolution to expand the channel to ensure that more features on the channel can be obtained from subsequent calculations. Considering that conventional self-attention cannot be used for such large-resolution image processing tasks, we take the approach proposed by Zamir et al. [29]: reshape the tensor before multiplying query (Q), key (K) and value (V); $Q\in \mathbb{R}^{HW\times C}$; $K\in \mathbb{R}^{C\times HW}$; $V\in \mathbb{R}^{HW\times C}$. After Q and K are multiplied, softmax operation is performed on the results to obtain the attention map of size $\mathbb{R}^{C\times C}$. The resulting attention map is multiplied by V to get the final attention feature. The advantage of transposing and reshaping before multiplication is that attention can be calculated based on channels. For this kind of large-resolution image, it is a very time-consuming and performance-consuming task to calculate attention map based on pixels. Overall, the MDTSA process is defined as:

$$\tilde{X} = W_\beta Attention(XW_1^Q W_3^Q, XW_1^K W_3^K, XW_1^V W_3^V) + X \tag{1}$$

where W_β is a learnable scaling parameter to control the size of the residual block, $W_1^{(\cdot)}$ is the 1×1 point-wise convolution and $W_3^{(\cdot)}$ is the 3×3 depth-wise convolution.

2.2 Dconv SimpleGate Feed-Forward Network

In the regular feed-forward network, two 1 × 1 convolutions and an activation function are included. In our study, we modified the conventional feed-forward network. After a 1 × 1 convolution, the same depth-wise convolution as MDTSA was introduced for local feature extraction. In the selection of activation function, we did not use nonlinear activation functions such as Gelu or Relu in this module, but use SimpleGate to replace the function of the activation function as the study by Chen et al. [6]. The feature map was directly divided into two parts according to the channel dimension, and then multiplied. In this way, the purpose of nonlinearity can also be achieved. The SimpleGate is defined as:

$$SimpleGate(X, Y) = X \cdot Y \tag{2}$$

where X and Y are two feature maps of the same size. $X \cdot Y$ can also achieve the purpose of nonlinearity of the activation function. And after SimpleGate, the number of channels of the feature will be reduced by 50%. This effect is difficult to achieve in common activation functions. In addition, we also add a 1 × 1 convolution to keep the number of channels unchanged.

2.3 Loss Function

We use the MSE loss function, which is widely used in the field of image restoration. MSE can well compare the pixel difference between the NDCT and model output, so that it can better guide the update of model parameters after gradient backward. It is defined as follows:

$$L_{mse} = \frac{1}{N}\sum_{i=1}^{N} \|D(x_i) - y_i\|^2 \tag{3}$$

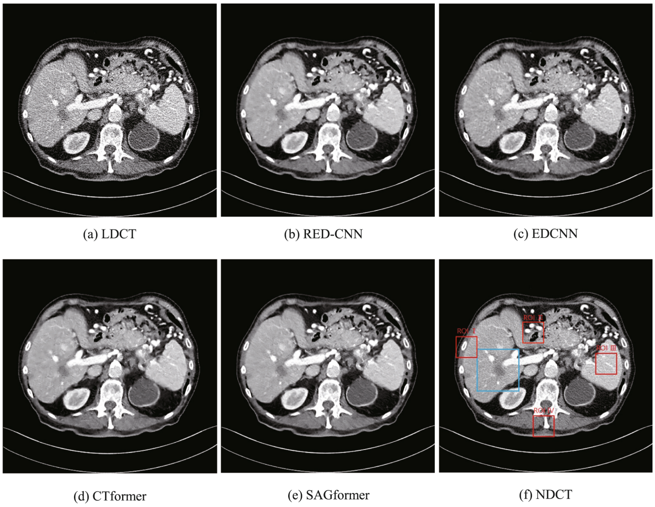
3 Experiments and Results

3.1 Experimental Settings

Datasets. In this work, we used the pubicly released dataset from 2016 NIH-AAPM-Mayo LDCT Grand Challenge [17]. The dataset was obtained by performing normal-dose abdominal CT scans on 10 anonymous patients, and then simulating quarter-dose images by adding Poisson noise to the projection data. We used the images from 9 patients to train the model and the images from 1 patient to evaluate the performance of the model. We randomly extracted eight image patches of 64 × 64 from each 512 × 512 image during training.

Implementation Details. In our work, we used Pytorch 1.11.0 to build our model, the training was performed on a NVIDIA RTX 3090Ti GPU. During training, we set the number of the epoch to 400, and used the Adam optimizer to minimize our MSE loss with an initial learning rate of 1e−5. All 3 × 3 depth-wise convolutions were implemented with stride 1 and padding 1, and 1 × 1 convolutions ere implemented with stride 1 and padding 0. During the test, we no longer divided the images into patches, but directly input 512 × 512 images into the model.

Baseline Models. Three baseline methods, i.e., RED-CNN [5], EDCNN [13], CTformer [23] were used to compared with our method. These methods are deep learning methods that perform well on LDCT denoising. But since RED-CNN and EDCNN did not provide a trained model, we retrained with the same dataset. CTformer used the same dataset as ours, and the authors released their trained model. So, we directly used the trained model for testing.


(a) LDCT (b) RED-CNN (c) EDCNN

(d) CTformer (e) SAGformer (f) NDCT

Fig. 2. Results from the different methods for comparison. The display window is $[-160, 240]$ HU. (a) a LDCT image; (b) RED-CNN; (c) EDCNN; (d) CTformer; (e) SAGformer. (f) a NDCT image. The red rectangle is the several defined ROIs. (Color figure online)

Evaluation Metrics. We used three common metrics to evaluate the methods, i.e., peak signal to noise ratio (PSNR), structural similarity index measure (SSIM) and root mean square error (RMSE) [23]. The three metrics can be combined to evaluate the denoising level, PSNR is mainly to measure the reconstruction quality. SSIM is used to evaluate the structural similarity between images, which can reflect the visual quality to a certain extent. RMSE is used to reflect the difference between corresponding pixels [14]. Among them, PSNR and SSIM are positively correlated with the final image quality, and RMSE is negatively correlated with the image quality.

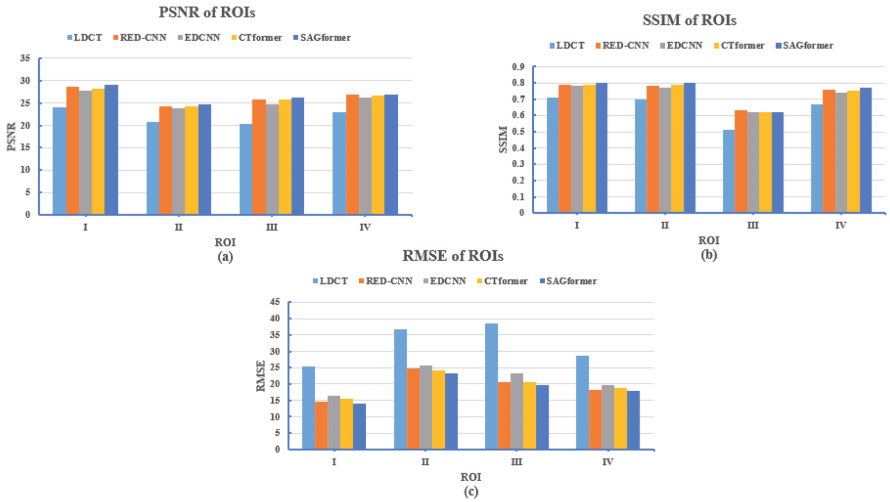
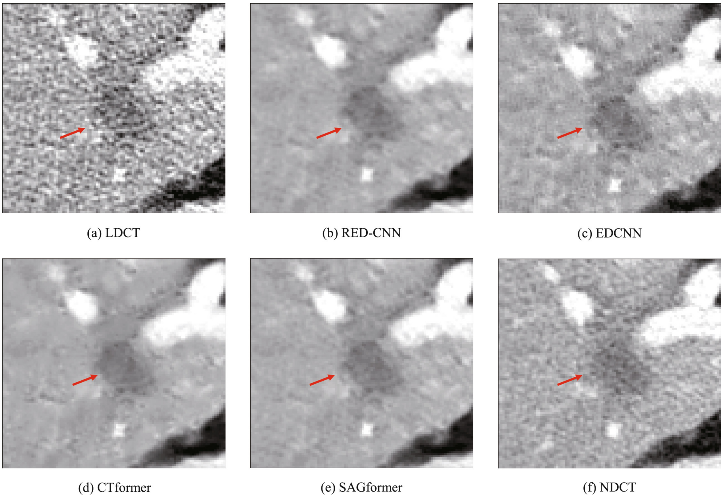


Fig. 3. The PSNR, SSIM and RMSE histogram of ROIs from Fig. 2 under different algorithms. (a) PSNR of ROIs; (b) SSIM of ROIs; (c) RMSE of ROIs.


(a) LDCT (b) RED-CNN (c) EDCNN

(d) CTformer (e) SAGformer (f) NDCT

Fig. 4. The amplified ROI images of different methods outputs in the blue rectangle marked from Fig. 2 (Color figure online).

3.2 Results

Quantitative Evaluation. First, we randomly selected four ROIs, as shown in the red rectangle of Fig. 2, to evaluate the reconstruction level of the local region. In order to make the results more intuitive, we display the values in the form of a histogram. As shown in Fig. 3, our results outperform other methods in PSNR, SSIM and RMSE. Second, we calculated the PSNR, SSIM and RMSE of each CT global image of patient L506, and then calculated the average values. Table 1 shows the overall quantitative results of all methods, and the best results are highlighted in bold. In terms of PSNR index comparison, SAGformer is 0.70 dB higher than the previous CNN-based method RED-CNN and 0.44 dB higher than the recent Transformer-based network CTformer. In the comparison of SSIM, our model is 0.005 higher than CTformer. Besides, the RMSE of the SAGformer is 0.45 lower than that of the CTformer.

Table 1. Quantitative results of different methods on L506

	PSNR↑	SSIM↑	RMSE↓
LDCT	29.2489	0.8759	14.2416
RED-CNN	32.8209	0.9092	9.3144
EDCNN	32.3036	0.9034	9.9121
CTformer	33.0793	0.9119	9.0726
SAGformer	**33.5281**	**0.9172**	**8.6172**

Qualitative Evaluation. Figure 2 shows the restoration effect of various models on a LDCT image. The abdominal CT slice was selected because the abdomen was the part with the most organs, so it can be seen clearly and intuitively that the noise reduction performance of our model was better than other models. Figure 4 shows the zoomed parts of the lesion position in the blue rectangle marked in Fig. 2. The lesion area pointed by the red arrow becomes more obvious after denoising, and the contour is also obvious.

3.3 Ablation Study

In this part, we verify the effectiveness of the introduced modules for LDCT denoising in the SAGformer through ablation experiments. The experimental results are shown in Table 2. We used multi-head transposed self-attention and feed-forward neural networks as benchmarks, which was termed as MTSA_FN. In the MTSA block, we used the normal 3 × 3 convolution to get QKV in self-attention, and in the FN network, we added the normal 3 × 3 convolution after the first 1 × 1 convolution. Although this model achieves better results than the CNN-based methods, the ordinary convolution operation results in a huge amount of parameters. Next, we replaced the 3 × 3 ordinary convolution with the depth-wise convolution, which resulted in the MDTSA_DFN model. The parameters of this model were reduced to 18.42 M, and the final PSNR, SSIM and RMSE are increased by 0.27, 0.004 and 0.27 respectively. Finally, we added SimpleGate to the feed-forward neural network module to form our final model (SAGformer), which once again improved PSNR, SSIM and RMSE. And the amount of parameters is reduced again because of the introduction of SimpleGate.

Table 2. Quantitative Results of Ablation Experiments on L506

	Params	PSNR↑	SSIM↑	RMSE↓
MTSA_FN	92.79M	32.9941	0.9114	9.1574
MDTSA_DFN	18.42M	33.2641	0.9159	8.8809
SAGformer	**16.64M**	**33.5281**	**0.9172**	**8.6172**

4 Conclusion

In brief, we designed a novel Transformer-based UNet network to successfully denoise low-dose CT images. Mainly through the modification of the self-attention calculation method, the tensor was reshaped and transposed to speed up the calculation speed. In addition, the use of depth-wise convolution and SimpleGate also greatly reduced the network parameters and improved the denoising performance. The experimental results show that the proposed model has great potential for structure preservation and lesion detection, and outperforms other models on both global and local regions after denoising. However, LDCT images still have some gaps compared to NDCT after denoising by our model, such as our images are over-smoothing. In addition, the very detailed textures have not been fully restored. In the future, we need to further optimize the SAGformer structure, and utilize the 3D CT image series to enhance the quality of low-dose CT images.

References

1. Brenner, D.J., Hall, E.J.: Computed tomographyan increasing source of radiation exposure. N. Engl. J. Med. **357**(22), 2277–2284 (2007)
2. Burger, H.C., Schuler, C., Harmeling, S.: Learning how to combine internal and external denoising methods. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 121–130. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40602-7_13
3. Cao, X., Yang, J., Gao, Y., Wang, Q., Shen, D.: Region-adaptive deformable registration of CT/MRI pelvic images via learning-based image synthesis. IEEE Trans. Image Process. **27**(7), 3500–3512 (2018)
4. Chen, C.H., et al.: Computer-aided diagnosis of endobronchial ultrasound images using convolutional neural network. Comput. Methods Programs Biomed. **177**, 175–182 (2019)
5. Chen, H., et al.: Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
6. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022)
7. Chen, Y., et al.: Improving abdomen tumor low-dose CT images using a fast dictionary learning based processing. Phys. Med. Biol. **58**(16), 5803 (2013)
8. De Man, B., Basu, S.: Distance-driven projection and backprojection in three dimensions. Phys. Med. Biol. **49**(11), 2463 (2004)

9. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Feruglio, P.F., Vinegoni, C., Gros, J., Sbarbati, A., Weissleder, R.: Block matching 3D random noise filtering for absorption optical projection tomography. Phys. Med. Biol. **55**(18), 5401 (2010)
11. Kang, D., et al.: Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm. In: Medical Imaging 2013: Image Processing, vol. 8669, pp. 671–676. SPIE (2013)
12. Lewitt, R.M.: Multidimensional digital image representations using generalized Kaiser-Bessel window functions. JOSA A **7**(10), 1834–1846 (1990)
13. Liang, T., Jin, Y., Li, Y., Wang, T.: EDCNN: edge enhancement-based densely connected network with compound loss for low-dose CT denoising. In: 2020 15th IEEE International Conference on Signal Processing (ICSP), vol. 1, pp. 193–198. IEEE (2020)
14. Luthra, A., Sulakhe, H., Mittal, T., Iyer, A., Yadav, S.: Eformer: edge enhancement based transformer for medical image denoising. arXiv preprint arXiv:2109.08044 (2021)
15. Ma, J., et al.: Low-dose computed tomography image restoration using previous normal-dose scan. Med. Phys. **38**(10), 5713–5731 (2011)
16. Mathews, J.P., Campbell, Q.P., Xu, H., Halleck, P.: A review of the application of X-ray computed tomography to the study of coal. Fuel **209**, 10–24 (2017)
17. McCollough, C.H., et al.: Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. Med. Phys. **44**(10), e339–e352 (2017)
18. Pan, X., Sidky, E.Y., Vannier, M.: Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction? Inverse Prob. **25**(12), 123009 (2009)
19. Ramani, S., Fessler, J.A.: A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction. IEEE Trans. Med. Imaging **31**(3), 677–688 (2011)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Shan, H., et al.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nat. Mach. Intell. **1**(6), 269–276 (2019)
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., Yu, H.: CTformer: convolution-free token2token dilated vision transformer for low-dose CT denoising. arXiv preprint arXiv:2202.13517 (2022)
24. Wang, D., Wu, Z., Yu, H.: TED-Net: convolution-free T2T vision transformer-based encoder-decoder dilation network for low-dose CT denoising. In: Lian, C., Cao, X., Rekik, I., Xu, X., Yan, P. (eds.) MLMI 2021. LNCS, vol. 12966, pp. 416–425. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87589-3_43
25. Whiting, B.R., Massoumzadeh, P., Earl, O.A., O'Sullivan, J.A., Snyder, D.L., Williamson, J.F.: Properties of preprocessed sinogram data in X-ray computed tomography. Med. Phys. **33**(9), 3290–3303 (2006)
26. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans. Med. Imaging **36**(12), 2536–2545 (2017)

27. Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D.: Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Trans. Biomed. Eng. **63**(7), 1505–1516 (2015)
28. Yang, Q., et al.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
29. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)
30. Zhang, W., et al.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage **108**, 214–224 (2015)
31. Zhang, Y., et al.: Clear: comprehensive learning enabled adversarial reconstruction for subtle structure enhanced low-dose CT imaging. IEEE Trans. Med. Imaging **40**(11), 3089–3101 (2021)
32. Zhang, Z., Yu, L., Liang, X., Zhao, W., Xing, L.: TransCT: dual-path transformer for low dose computed tomography. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 55–64. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_6