Low-Dose CT Denoising via Sinogram Inner-Structure Transformer

Liutao Yang, Zhongnian Li, Rongjun Ge, Junyong Zhao, and Daoqiang Zhang, Senior Member, IEEE

Abstract—Low-Dose Computed Tomography (LDCT) technique, which reduces the radiation harm to human bodies, is now attracting increasing interest in the medical imaging field. As the image quality is degraded by low dose radiation, LDCT exams require specialized reconstruction methods or denoising algorithms. However, most of the recent effective methods overlook the inner-structure of the original projection data (sinogram) which limits their denoising ability. The inner-structure of the sinogram represents special characteristics of the data in the sinogram domain. By maintaining this structure while denoising, the noise can be obviously restrained. Therefore, we propose an LDCT denoising network namely Sinogram Inner-Structure Transformer (SIST) to reduce the noise by utilizing the inner-structure in the sinogram domain. Specifically, we study the CT imaging mechanism and statistical characteristics of sinogram to design the sinogram inner-structure loss including the global and local inner-structure for restoring high-quality CT images. Besides, we propose a sinogram transformer module to better extract sinogram features. The transformer architecture using a self-attention mechanism can exploit interrelations between projections of different view angles, which achieves an outstanding performance in sinogram denoising. Furthermore, in order to improve the performance in the image domain, we propose the image reconstruction module to complementarily denoise both in the sinogram and image domain.

Index Terms—Low-dose CT, deep learning, vision transformer, sinogram inner-structure.

I. INTRODUCTION

COMPUTED Tomography (CT) is an important medical imaging modality for medical diagnosis. However, exposure of patients to the X-ray radiation of CT examinations would increase the risk of cancer. Therefore, the Low-Dose Computed Tomography (LDCT) technique is in urgent need in the clinic. The most common way to reduce the radiation dose in CT scans is to reduce the X-ray tube current (or voltage). As a consequence, images that are reconstructed from this current (or voltage) always suffer from severe noise.

In order to effectively improve the imaging quality of LDCT, lots of methods are developed in the past decades. They can be roughly divided into two categories: 1) sinogram domain reconstruction and 2) image domain post-processing. Sinogram domain methods focus on the original projection data. They either use filters [1], [2] on the projection data to smooth the sinogram or use iterative reconstruction [3], [4], [5], [6], [7], [8], [9] based on priors. However, filtration methods always result in spatial resolution loss in the reconstructed image. Iterative reconstruction methods are time-consuming. Since the deep learning methods [10], [11], [12], [13], [14] achieve great progress in medical image segmentation, reconstruction, and analysis, image domain post-processing [15], [16], [17], [18], [19], [20] methods of LDCT denoising attain a state-of-the-art performance by utilizing the deep learning framework. As pure image domain methods overlook the origin projection data, several methods [21], [22] add the sinogram domain information into the deep learning denoising network to further improve the performance. However, these methods, without consideration of the inner-structure of original projection data, have limitations on the denoising ability. Due to the physical mechanism of CT imaging, the sinogram contains a special inner-structure compared to the natural image. For example, as shown in Fig.1 (a), conjugate sampling pairs during CT scans receive X-rays with the same path so that corresponding points in the sinogram should obtain the same value. For each point of the sinogram, there is a corresponding conjugate point resulting in a conjugate structure over the sinogram domain. Considering that noise appears randomly (e.g., Poisson Distribution) in the whole imaging procedure, the conjugate pairs obtaining different noise would evidently break this structure. By maintaining this sinogram inner-structure, we can effectively restrain the noise and improve the image quality. Therefore, exploring the inner-structure of sinogram is of great importance for sinogram domain denoising.

In this paper, we study the CT imaging mechanism and statistical characteristics to propose the Sinogram Inner-Structure Loss (SISL). As we illustrated above, the
We can see the second-order derivatives are of high sparsity. By maintaining these inner-structures while sinogram denoising, we can significantly improve the image quality.

The sinogram inner-structure represents the typical data characteristics in the sinogram domain. And the noise, which appears randomly while imaging, would break this structure. Designing a loss function based on sinogram inner-structure for network training in the sinogram domain can effectively maintain this structure and restrain noise. The inner-structure loss is designed both at the global and local levels. The global inner-structure utilizes conjugate sampling pairs in CT scans. As Fig. 1 (a) shows, this relation constrains conjugate projection pairs in the sinogram to obtain the same value. Since the correlation exists in the whole sinogram, it can help maintain the global inner-structure. The local inner-structure considers the sparsity of second-order derivatives in the sinogram domain. As illustrated in [23], the sinogram can be described as a piecewise linear configuration. So, we can obtain that second-order derivatives of the sinogram will be very sparse.

Fig. 1 (b) shows horizontal second-order derivatives of the sinogram. This feature is calculated between adjacent points in the sinogram which can help maintain the local inner-structure. In general, the proposed Sinogram Inner-Structure Loss is essentially the regularization of sinogram domain which compel the sinogram data to consistent with the data characteristic caused by CT imaging mechanism. Thus, the denoised sinogram can be limited to the feasible data space which can significantly improve the effectiveness of sinogram denoising.

For network design, most methods use a CNN-based backbone in both the image and sinogram domain. However, we argue that the transformer architecture is more applicable in the sinogram domain. The transformer shows outstanding performance in NLP field by utilizing the multi-head self-attention to extract interrelation between sequence data. According to the CT imaging mechanism, each row of the sinogram is the projection at a certain view angle. By regarding the sinogram as sequences of projection, the self-attention mechanism in the transformer can extract relations between projections under different view angles, which is hard to achieve in a CNN architecture. Furthermore, since pure sinogram domain denoising can lead to artifacts in reconstructed images, we use an image reconstruction module to transfer the sinogram noise into the image domain and apply image domain denoising in one unified network. Thus, the image domain loss can be back-propagated into the sinogram domain for complementary optimization.

Contributions of this paper can be summarized as follows:

- We propose the global inner-structure loss which utilizes conjugate sampling pairs in CT scans and local inner-structure loss which considers the second-order sparsity of sinograms. These losses can help to better use the sinogram information and improve the denoising performance.
- We propose a Sinogram Inner-Structure Transformer (SIST) for LDCT denoising. This network, taking each view angle of sinogram as input, can more effectively extract the structure feature than CNN-based networks.
- We propose an image reconstruction module to transfer the sinogram noise into the image domain and apply image domain denoising in one unified network. Thus, the image domain loss can be back-propagated into the sinogram domain for complementary optimization.

The rest of the paper is organized as follows: In Related Works, we briefly introduce the LDCT reconstruction/denoising methods and transformers used in computer vision. Then, in the Proposed Method, we give the detail of our proposed Sinogram Inner-Structure Transformer. In Experiments, we compare the proposed method with state of the art and conduct several ablation studies. Finally, we give a conclusion in the last section.

II. RELATED WORKS

A. Low Dose CT Reconstruction and Denoising

Due to the importance of LDCT, lots of model-based image reconstruction methods have been proposed during the past decades. These methods, which are based on priors of CT images, system model, or measurement statistical model, iteratively find the optimal image from the projection data. Ma et al. [24] proposed the previous normal-dose scan induced nonlocal means (ndiNLM) method to utilize the normal-dose image to enable low dose CT image reconstruction. Dose reduction using prior image constrained compressed sensing (DRPICCS) [25] is proposed to reduce image...
noise using compressed sensing. Recently proposed PWLS-ULTRA [26] exploits loss based on an efficient Union of Learned TRAnsforms which is pre-learned from numerous image patches extracted from a dataset of CT images. By using the learning-based method in iterative reconstruction, this method can achieve a good performance.

Different from the model-based image reconstruction methods focus on the reconstruction phase, deep learning based CT image denoising tries to remove the artifact and noise in reconstructed images. Inspired by the SRCNN [27], Chen et al. [19] proposed a residual encoder-decoder CNN (REDCNN) for LDCT. With the development of GAN, Wolterink et al. [28] propose a method to train a generative CNN jointly with an adversarial CNN to estimate NDCT images from LDCT images and hence reduce noise. Then Shan et al. [20] proposed a modularized adaptive processing neural network (MAP-NN), which performs an end-to-process mapping with a modularized neural network to optimize the denoising depth in a task-specific fashion. More recently works start to focus on the denoising on both the sinogram and image domain. Yin et al. [21] proposed a domain progressive 3D residual convolutional network (DP-ResNet) to denoising both in sinogram and image domain. This method train sinogram denoising CNN and image denoising CNN separately. Then they reconstruct the denoised sinogram into images and feed into the image domain CNN for further denoising. Hu et al. [29] propose the hybrid-domain neural network (HDNet) which decomposes the CT reconstruction problem into two stages to reduce the learning difficulty of the entire network. Wang et al. [30] propose a deep reconstruction network to achieve end-to-end training for both sinogram and image domains. By incorporating the FBP/FDK algorithm into the network, training a fully connected layer to convert the sinograms to CT images is avoided and the number of weights in network is decreased.

However, all those deep learning methods either only consider the image domain information or overlook the inner-structure of original projection data. Thus, our proposed method which uses the sinogram inner-structure to assist the LDCT denoising can considerably improve the performance.

B. Transformers in Computer Vision

The transformer is first proposed [31] for natural language processing. Due to the competitive performance in many tasks, it soon became a very important architecture used in Natural Language Processing (NLP). For example, GPT [32], [33] pre-trains in an auto-regressive way that predicts the next word in huge text datasets. BERT [34] utilizes the transformer to predict a masking word based on context.

The success of transformer in NLP encourage researchers to attempt its applications in computer vision task. Dosovitskiy et al. proposed the Vison Transformer (ViT) [35] to directly split the image into patches and provide the sequence of linear embeddings. This sample but effective architecture attains excellent results in image classification. Wu et al. [36] represent images as semantic visual tokens and run transformers to densely model token relationships. This method achieves impressive results both in image classification and semantic segmentation. Hassani et al. [37] propose the Compact Transformer using convolutional tokenization in transformer head and achieves state-of-the-art performance in small datasets classification. Pre-trained image processing transformer (PIT) [38] introduces the transformer into the low-level computer vision task (e.g., denoising, super-resolution, and deraining). With convolutional tokenization and pretrained on large datasets, this method also achieves the state of the art.

In all these works above, we observe that the key point for vision transformer is the tokenization method. In another word, how to transfer the 2D structure image into the 1D structure as input can largely impact the performance of vision transformers. As the sinogram is essentially 1D projection data, the transformer can better extract the information than a CNN architecture. So in this paper, we design the sinogarm transformer for LDCT denoising.

III. METHOD

In this section, we introduce each part of the proposed Sinogram Inner-Structure Transformer (SIST). We first introduce the sinogram transformer module for sinogram domain denoising. Then we give the details of the Sinogram Inner-Structure Loss (SISL). Besides, we introduce the image reconstruction module which converts sinogarm noise into image domain for complementarily denoising both in sinogarm and image domain. Finally, we give the total loss function used for training. The overall architecture is shown in Fig.2.

A. Preliminaries

The LDCT denoising task in this paper is to recover the NDCT sinogram $S_{nd} \in \mathbb{R}^{P \times D}$ and image $I_{nd} \in \mathbb{R}^{W \times H}$ from the LDCT sinogram $S_{ld} \in \mathbb{R}^{P \times D}$ and image $I_{ld} \in \mathbb{R}^{W \times H}$, where $P$ is the number of projection views, $D$ is the number of detectors, $W$ and $H$ are the width and height of images. $s_{ld}^i \in \mathbb{R}^D, i = \{1, 2, \ldots, P\}$ is projections in $i_{th}$ view angle of $S_{ld}$. Note that $I_{nd}$ and $I_{ld}$ can be reconstructed from the sinogram $I_{nd}$ and $I_{ld}$ using reconstruction algorithm such as Filtered Back-Projection (FBP). $\hat{S}$, $\hat{I}$ and $I_{noise}$ are the generated sinogarm, image and image noise respectively.

B. Sinogarm Transformer Module

The transformer architecture can more effectively extract the sinogarm feature from different view angles than CNN-based networks due to the multi-head self-attention mechanism. So, we propose the sinogarm transformer module for sinogarm domain denoising. In traditional vision transformer structures, image is cut into patches as the input token of transformer considering the spatial relation of nature images. However, the sinogarm is essentially the projection data at different imaging views, cutting patches would evidently break the consistency of data at the same view. Besides, because each view represents one measurement of imaging target, modeling self-attention relations between views can benefit the network to better learn the general feature under different views which is crucial for sinogarm denoising. Thus, we design
the sinogram transformer to apply the per-angle learning for sinogram denoising. Same as the typical transformer design, the sinogram transformer module consists of three parts: head, transformer encoder, and tail. The detail of each part are shown as follows:

1) **Head**: Since tokenization is the key component of the performance, we split each row of the sinogram as a set of sequence data. Then we use an MLP to embed the sinogram into the input:

\[ H = \text{ReLU}(\text{LN}((s_{ld}^1, s_{ld}^2, \ldots, s_{ld}^P))_{\times n}) \]  

where LN is the linear layer convert sinogram \( s_{ld} \) with dimension \( D \) into embeddings dimension \( D' \). ReLU is the activate function. \( H \in \mathbb{R}^{n \times D'} \) is the extracted feature as the input of transformer. The Multilayer Perceptron (MLP) consists of this block repeated for \( n \) times.

2) **Transformer Encoder**: In this part, we follow the original Vision Transformer design. Details can refer to [31]. We add the learnable position encodings \( e_i \in \mathbb{R}^{D'} \) to each projection of the transformer input \( h_i \) to maintain the position information [35], [39]. Then \( e_i + h_i \) is input into the transformer encoder directly. The transformer encoder consists of Multi-head Self-Attention (MSA), which allows the model to jointly attend to information from different representation subspaces at different positions [31], and Multilayer Perceptron (MLP) that is the same as Eq.1. The entire structure is shown as follows:

\[
\begin{align*}
Z_0 &= [e_1 + h_1; e_2 + h_2; \ldots; e_P + h_P], \\
Q_i &= K_i = V_i = \text{LN}(Z_{i-1}), \\
Z_i' &= \text{MSA}(Q, K, V) + Z_{i-1}, \quad i = 1, \ldots, l \\
Z_i &= \text{MLP}(Z_i') + Z_i' 
\end{align*}
\]  

where \( l \) is the number of layers in the transformer encoder. \( LN \) is the linear layer.

3) **Tail**: In the tail, We add a skip connection from the head to maintain the low-level feature. We use an MLP to recover the output of transformer encoder from dimension \( D' \) into \( D \):

\[ S' = \text{ReLU}(\text{LN}(Z_l + H))_{\times n} \]  

Then the output \( S' \) is restored to the shape \( \mathbb{R}^{P \times D} \). Finally, residual blocks are used for the final refinement of the sinogram:

\[ \hat{S} = \text{Residual}(S') \]  

After this refinement, we get the sinogram domain output \( \hat{S} \) for loss computing and next stage reconstruction.

### C. Sinogram Inner-Structure Loss

Due to the physical mechanism of CT imaging, the sinogram contains a special inner-structure compared to the natural image. As illustrated in Section I, the sinogram inner-structure represents the typical data characteristics in the sinogram domain. The noise, which appears randomly while imaging, would break this structure. Therefore, exploring the inner-structure of sinogram is of great importance for sinogram domain denoising. In this paper, we explore the inner-structure in both global and local levels, which are shown in Fig. 1. The global inner-structure utilizes conjugate sampling pair in CT scans. This relation constrains the corresponding points to obtain the same value. Since this correlation exists in the whole sinogram, it can be used for maintaining the global inner-structure. The local inner-structure uses the second-order derivatives sparsity of sinogram to establish the correlation between adjacent points. This feature indicates the correlation between the sinogram’s adjacent points, which helps maintain the local inner-structure.

For deep learning network training, we design the Sinogram Inner-Structure Loss (SISL) based on the both global and local inner-structure.

The SISL is designed in two levels: the global inner-structure loss \( L_C \) and local inner-structure loss \( L_S \):

\[ SISL = L_C + L_S \]  

The details of each loss are shown as follows:
projections in a 2D space with rotating the CT gantry to view angle of $\beta + \pi + 2\gamma$, where $\gamma$ is the detector angle of one detector. Then Eq. 6 can be discretely rewritten as:

$$S(i, j) = S(i + \text{int}(\frac{\pi + 2\gamma_d P}{2\pi}), -j)$$  \hspace{1cm} (7)

Thus, for each point $S(i, j)$ in $S$, its conjugate sampling should obtain the same value. Based on this fact, we can define the conjugate sinogram $\hat{S}_C$ of $\hat{S}$ using Eq. 7. Since the noise appears randomly in the imaging procedure, the conjugate relation would be broken in low-dose scans. As a consequence, the conjugate sinogram $\hat{S}_C$ would be different from $\hat{S}$ in some degree (depending on the noise level). By formulating a loss function based on the distance between $\hat{S}_C$ and $\hat{S}$, we can maintain the inner-structure and restrain noise while training. The global inner-structure loss can be formulated as:

$$L_C = \|\hat{S} - \hat{S}_C\|$$  \hspace{1cm} (8)

This loss calculates the Euclidean distance between $\hat{S}_C$ and $\hat{S}$. By minimizing this loss while network training, the noise can be effectively reduced.

2) Local Inner-Structure Loss: Fig.1 (b) gives the visualization of horizontal second-order derivatives of the sinogam. As illustrated in [23], the sinogram can be described as a piecewise linear configuration so that we can obtain that second-order derivatives of sinogram will be very sparse. To utilize this inner-structure in our deep learning framework, one feasible way is to use the Hessian penalty [40], [41] to introduce the sparsity of second-order derivatives. However, if the second-order derivative is over-sparsified while training, it would result in over-smooth in images. Since it’s hard to find an appropriate hyper-parameter to control the sparsity, we rather use the loss between the low-dose sinogram and ground truth (i.e., normal-dose sinogram). So, we define the local inner-structure loss of sinogram based on second-order derivatives:

$$L_S = \sum_{i,j} \sqrt{D_{ii}^2 + D_{jj}^2 + D_{ij}^2 + D_{ji}^2},$$

$$D_{ii} = \frac{\partial^2 S(i, j)}{\partial i^2} - \frac{\partial^2 S_{nd}(i, j)}{\partial i^2},$$

$$D_{jj} = \frac{\partial^2 S(i, j)}{\partial j^2} - \frac{\partial^2 S_{nd}(i, j)}{\partial j^2},$$

$$D_{ij} = \frac{\partial^2 S(i, j)}{\partial i \partial j} - \frac{\partial^2 S_{nd}(i, j)}{\partial i \partial j},$$

$$D_{ji} = \frac{\partial^2 S(i, j)}{\partial j \partial i} - \frac{\partial^2 S_{nd}(i, j)}{\partial j \partial i}$$

where $D_{ii}, D_{jj}, D_{ij}, D_{ji}$ are differences of second-order derivatives between output $\hat{S}$ and ground truth $S_{nd}$. 

---

**Image Reconstruction Module**

![Image Reconstruction Module](image)

Fig. 3. The architecture of the image reconstruction module. The sinogram domain noise is converted into the image domain. This module uses $S_{ld}$ minus $\hat{S}$ to generate the sinogram noise $S_{noise}$ and transfer it into image domain $I_{noise}$. Using $I_{ld}$ minus $\hat{I}_{noise}$, we can get the coarse denoised image. After the final image refinement, we can get the denoised image $\hat{I}$.

Fig. 4. Examples of LDCT images in Simulated Dataset under different dose levels. The LDCT images are generated using the noise inserting method in [42].
D. Image Reconstruction Module

The architecture above achieves effective sinogram domain denoising. We propose the image reconstruction module to apply image domain denoising complementarily. As shown in Fig.3, the sinogram domain noise is converted into the image domain. By removing the estimated noise from the LDCT image, a coarse denoised image is generated as the input of the later image denoising network.

Since Anirudh et al. [43] has successfully achieved domain transferring in a limited view CT restoring, we use a similar structure CNN as [43] but removing the sinogram complementing module. Instead of directly reconstructing the sinogram to the image, we first convert the sinogram noise into the image domain. Then we remove the estimated noise from the LDCT image, a coarse denoised image is generated as the input of the later image denoising network.

By using this design, the LDCT images can be input into the denoising network to add extra image domain information and guide the domain transfer. The learned reconstruction is actually between the sinogram noise and image noise:

$$\hat{I}_{\text{noise}} = \text{Residual}(\text{Conv1d}(S_{ld} - \hat{S}))$$  \hspace{1cm} (10)

where $\hat{I}_{\text{noise}}$ is the estimated noise between LDCT and NDCT images, $\text{Conv1d}$ is a 1D CNN embeds the sinogram noise into a latent space [43]. Then we use several residual blocks to transfer sinogram embeds into the image domain. finally, we remove the estimated image noise $\hat{I}_{\text{noise}}$ from LDCT images, and then input it into the next stage:

$$\hat{I} = UNet(I_{ld} - \hat{I}_{\text{noise}})$$  \hspace{1cm} (11)

E. Total Loss Function

In our proposed framework, The total loss function consists of four parts: sinogram denoising loss, sinogram inner-structure loss, image reconstruction loss, and image denoising loss. It can be formulated as:

$$L = \| \hat{S} - S_{nd} \|_{L1} + \text{SISL} + \| \hat{I} - I_{nd} \|_{L1} + \| \hat{I}_{\text{noise}} - (I_{ld} - I_{nd}) \|_{L1}$$  \hspace{1cm} (12)

where $\hat{I}$ is the final denoised image and $UNet$ is CNNs refer to [44]. This stage can further refine the image to achieve a better denoising performance.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on two datasets to evaluate the performance of our proposed method:

1) Low-Dose CT Image and Projection Dataset: The Low-Dose CT Image and Projection Dataset (LDCT Dataset) [45] consists of CT patient scans from three common exam types: non-contrast head CT scans acquired for acute cognitive or motor deficit, low-dose non-contrast chest scans acquired to screen high-risk patients for pulmonary nodules, and...
contrast-enhanced CT scans of the abdomen acquired to look for metastatic liver lesions. It contains both the projection and image data from 150 clinically performed patient CT exams SOMATOM Definition CT system. The scanning parameters are as follows: 576 projection views evenly spanning a 360° circular orbit were collected using a z-flying focal spot, resulting in 1152 projections within one rotation. The number of detector bins in each projection is 736. All these CT scans are conducted by a Fan-beam geometry. The width of detector is 1.28 mm, the in-plane distance between focal center and isocenter is 600.45 mm, and distance between isocenter and detector element is 485.15 mm. All 25141 samples have low-dose and normal-dose pairs. We randomly split 120 patients for training, and the remains are for testing.

2) Simulated Dataset: The Simulated Dataset consists of CT images from spine CT exams. All 46 patients are collected from Qilu Hospital of Shandong University. Since only NDCT images are available in this dataset, we use the ASTRA tomography Toolbox [46] to generate the projection data. We use the same geometry settings in LDCT Dataset. In order to simulate the LDCT exams, the noise is inserted using a previously validated photon-counting model that incorporates
the effect of the bowtie filter, automatic exposure control, and electronic noise [42]. In this dataset, we generate the LDCT in different dose levels: 5%, 10%, and 20% of the normal dose. The noise inserting function is shown as follows:

\[ P_B = P_A + \sqrt{\frac{1 - a \exp(P_A)}{N_0 A}} \left( 1 + \frac{a N_e \exp(P_A)}{N_0 A} \right) x \]  

(13)

where \( P_A \) is the logarithm-transformed projection data of the normal dose, \( N_0 A \) is the incident number of photons, \( P_B \) is the simulated low dose projection data. \( a \) indicates the dose levels of the simulated scans, \( x \) is a random variable that follows a standard normal distribution, and \( N_e \) is the noise-equivalent quanta of electronic noise. According to [42], \( N_0 A \) is set to \( 10^5 \). \( N_e \) is set to 10. In our experiments, \( a \) is set to 0.05, 0.1, and 0.2 for dose-levels: 5%, 10%, and 20% respectively. The training set contains 35 patients, and the remains are for testing.

B. Implementation Details

In the sinogram transformer module, the head uses two linear layers and converts the sinogram into dimension 1024. In the tail, the residual block consists of two residual convolutional layers. The MSA in the transformer encoder has 6 heads and 6 layers. The network is implemented by PyTorch with python 3.6 on 4 NVIDIA RTX 2080Ti. We used ADAM with momentum for optimization and set the initial learning rate to be \( 10^{-5} \), the first-order momentum to be 0.9, and the second momentum to be 0.999. A stepLR is used with a 0.7 decay rate every 10 epochs.

For image quality evaluation, three metrics are used in our experiments: peak signal to noise ratio (PSNR), structural similarity index measure (SSIM) and root mean square error (RMSE).

C. Comparison With Other Methods

In order to verify the performance of the proposed method, we conduct the comparison experiment with several state-of-the-art methods, i.e., PWLS-EP, PWLS-ULTRA, RED-CNN, and DP-ResNet. Both the PWLS-EP [26] and PWLS-ULTRA [26] are iterative reconstruction methods that utilize the low-dose sinogram. RED-CNN [19] and DP-ResNet [21] are CNN-based methods. RED-CNN only denoise in the image domain and DP-ResNet uses both sinograms and images.

1) LDCT Dataset: We first conduct the experiments on the Low-Dose CT Image and Projection Dataset. Tab.I gives quantitative results of all methods. Note that FBP in Tab.I are images directly reconstructed from a low-dose sinogram using FBP. These results show that deep learning methods outperform iterative reconstruction methods thanks to the delicate designed CNN architectures. Our method achieves the best performance in all three metrics by further utilizing the sinogram inner-structure information. Visual results are shown in Fig.5. It can be observed that our proposed method outperforms others. Fig. 6 shows examples of clinical diagnostic reconstructions in which the red circle indicates a liver cyst. In the LDCT image, this cyst is blurred by the noise. However, the proposed method can restore this area more successfully than other comparison methods.

To analyze the computational cost of the proposed method, we compare the runtime between different methods including both iterative reconstruction and deep learning methods at test phase. Tab. I shows the reconstruction performance and runtime of different methods. Note that the iteration times are set to 10 to balance the computational cost and performance. As Tab.I shows, deep learning methods have great advantages in reconstruction speed comparing to iterative reconstruction. The dual-domain methods (i.e., SIST and DP-ResNet) cost more runtime than image domain methods (i.e., RED-CNN) to gain better performance. Our propose SIST achieve the

| Method     | PSNR(↑) | SSIM(↑) | RMSE(↓) | Runtime |
|------------|---------|---------|---------|---------|
| FBP        | 33.42±3.78 | 0.752±0.078 | 0.0064±0.0077 | 0.44s   |
| PWLS-EP    | 36.76±1.45 | 0.801±0.026 | 0.0043±0.0021 | 103.45s |
| PWLS-ULTRA | 38.98±1.33 | 0.902±0.022 | 0.00337±0.0011 | 212.64s |
| RED-CNN    | 40.31±1.43 | 0.908±0.013 | 0.00124±0.0019 | 0.67s   |
| DP-ResNet  | 40.92±1.23 | 0.914±0.011 | 0.00269±0.0012 | 3.12s   |
| SIST(Our)  | 41.80±1.37 | 0.916±0.012 | 0.00246±0.0009 | 0.71s   |

| Method     | PSNR(↑) | SSIM(↑) | RMSE(↓) |
|------------|---------|---------|---------|
| FBP        | 28.22±5.87 | 0.625±0.150 | 0.0422±0.0067 |
| PWLS-EP    | 36.88±1.91 | 0.843±0.018 | 0.0091±0.0021 |
| PWLS-ULTRA | 37.22±1.01 | 0.893±0.011 | 0.0087±0.0012 |
| RED-CNN    | 40.67±1.22 | 0.948±0.008 | 0.0059±0.0012 |
| DP-ResNet  | 41.43±1.65 | 0.952±0.014 | 0.0054±0.0015 |
| SIST(Our)  | 42.62±1.14 | 0.973±0.009 | 0.0047±0.0007 |

**TABLE I**

Experimental Results for Different Methods on LDCT Dataset. Best results are highlighted and second results are underlined. ↓ (↑) means the lower (higher) the better.

**TABLE II**

Experimental Results for Different Methods on Simulated Dataset. Best results are highlighted and second results are underlined. ↓ (↑) means the lower (higher) the better.

The training set contains 35 patients, and the remains are for testing.
best performance with acceptable runtime cost using a unified network design.

We further perform the t-test on the PSNR results achieved by different methods. Our proposed SIST shows the significant improvement \((p < 0.05)\) over DP-ResNet \((p = 1.686e-05)\), RED-CNN \((p = 3.577e-05)\), PWLS-ULTRA \((p = 1.328e-08)\), PWLS-EP \((p = 7.969e-11)\).

2) Simulated Dataset: To further evaluate the performance of the proposed method on different dose levels, we generate the projection data and insert noise follow [42]. We evaluate the performance on three dose levels: 20%, 10%, and 5%. The quantitative results are shown in Tab.II. From the table, we can observe that our methods achieve the best results in all dose levels: 20%, 10%, and 5%. Same as the LDCT
dataset, the deep learning methods show superior performance that iterative reconstruction methods. Moreover, our method still has great performance at the very low dose situation (5% dose), which even over the path the second method in 10% dose. Fig.7, Fig.8, and Fig.9 show the visual results for dose levels 20%, 10%, and 5%. From these visual results, we can observe that our method can preserve more details while denoising.

Fig. 11. The influence of loss weights on performances. In each experiment, we set the term of interest to 0.1, 0.5, 1, and 5 respectively while keeping the rest to 1.

D. Ablation Study

To further explore the effectiveness of different components, we conduct the ablation study in this section.

1) Effectiveness in Image Domain: Compare to typical deep learning methods, the proposed framework focuses on the inner-structure of the sinogram to improve the image domain denoising quality. So, one important question is that how much does the inner-structure impact the image quality? Tab.III shows results of using different components. We first remove all sinogram domain components and use the rest part (U-Net) to train and test only in the image domain. Without any specific design, the PSNR achieves 38.09 dB of pure U-Net CNNs. Since SIST trains dual-domain network end-to-end, midterm losses (i.e., sinogram loss and noise loss) are important for network convergence. As the results show, the performance of SIST gains little improvement without midterm losses. Actually, without midterm losses, the sinogram transformer module and image reconstruction module would easily reach a local minimum where they cannot accomplish the tasks, which gains very little improvements on performance. After we add the sinogram loss and image loss, the performance improved significantly in all three metrics. To illustrate the effectiveness of transformer structure, we also replace the sinogram transformer module with a fully connected network (FC+U-Net) using per-angle data to compare the performance. The results show that sinogram transformer module can achieve higher performance when using the same loss terms.

To further utilize the inner-structure of the sinogram, the global inner-structure loss (Lc) is added to utilize conjugate projection pairs in the sinogram. This loss helps improve the performance as shown in Tab.III. At last, we add the local inner-structure loss (LS) to maintain the second-order sparsity of the sinogram, and the performance is further improved.

2) Effectiveness in Sinogram Domain: In this part, we verify the effectiveness of our method in improving the sinogram denoising. For comparison, we use the U-Net and sinogram transformer module to train with low-dose/normal-dose sinogram pairs. Since we care more about the quality of reconstructed images, we evaluate sinograms by applying FBP reconstruction. Note that our proposed method use end-to-end training to directly get the reconstructed image. For a fair comparison, we still apply FBP on the intermediate sinogram output. Fig.10 shows the example of the results. As we can observe, even though all the denoised sinograms are of high similarity to the ground truth, there still are obvious artifacts in reconstructed images. Compare to the CNN-based U-Net, the sinogram transformer can better extract the structure information in the sinogram and improve the quality. By further adding the inner-structure loss of sinogram, artifacts in images are considerably reduced.

3) The Influence of Loss Weights: In this section, we analyze the influence of loss weights on performance. For a fair comparison, we set the term of interest to 0.1, 0.5, 1, and 5 respectively for each loss term while keeping the rest to 1. In Fig.11, reducing the weight of each term would degrade performance, which shows that all these terms have important influences on network performance. Since the trend of each term is different, the importance of each term shows some difference. For sinogram loss, noise loss, and SISL (sinogram inner-structure loss), the network achieves high performance when weights exceed 0.5. There are no significant improvements when continuing to increase the weights (the PSNR even shows slight declines when weights exceed 1). The image loss shows a higher dependency on PNSR since it is more relevant to the computation of PNSR. We also report the SSIM for the
same weight setup to evaluate the influence of loss weights on image structure. As results show, the SISL is more relevant to the structural similarity. Higher weights of image loss would result in obvious declines in SSIM which would lead to blurry details. As exhaustive searching of the weights requires a large amount of computational resources, we set all terms equally to balance performance between pixel level (PSNR) and structure level (SSIM), which still shows a promising performance.

E. Noise and Spatial Resolution Evaluation

For denoising performance evaluation, we compute the average Noise Power Spectrum (NPS) for the proposed method and the Dp-ResNet which has the best performance among other comparison methods. The NPS is calculated following [30] and the two-dimensional NPS is circularly averaged [47] into one-dimensional profile in Fig. 12.

As shown in [42], LDCT images with lower dose have a higher frequency of noise, which indicates that low-dose noises are mostly in a high-frequency range. In Fig. 12, both methods can effectively reduce high-frequency noise. Actually, most LDCT reconstruction algorithms have the trend to remove the high-frequency noise resulting in NPS peak frequency shift to a lower range. However, because most objects are low-frequency dominated, too much low-frequency noise (e.g., Dp-ResNet) could have a potentially deleterious effect on the signal-to-noise ratio and object detectability [47]. Thus, this is a trade-off in LDCT reconstruction which suggest that reconstruction methods need to keep a balance between high-frequency and low-frequency. As shown in Fig. 12, the proposed SIST not only reduces the low-dose noise in the high-frequency range but also shows a better performance in the low-frequency range than the second method (i.e., Dp-ResNet). This feature indicates the proposed SIST could have better performance in lesion detectability and clinical practice.

To test the spatial resolution of the proposed method, the modulation transfer function (MTF) is calculated using the radial-stripe image in [30]. We still compare the proposed method with Dp-ResNet on both dose levels and image object contrast levels. As shown in Fig. 13, in a high contrast level (i.e., the HU value of the radial-stripe image is in [0, 200]), both methods have good spatial resolution performance in all three dose levels. In lower contrast levels (HU value in [0, 100] and [0, 50]) the performance of Dp-ResNet is obviously degraded especially in lower doses such as 5%). In these situations, our proposed method shows better spatial resolution in all dose levels. Thus, our method has a better performance on image details.

V. CONCLUSION

In this paper, we introduce the sinogram inner-structure transformer which utilizes the sinogram domain information to assist the LDCT denoising in the image domain. In order to improve the sinogram quality while training, we introduce the inner-structure loss to maintain the special structure of the sinogram. The inner-structure loss considers both the global and local inner-structure. Thus, the proposed SIST achieves a great performance in the LDCT denoising task. Experiments on two datasets also show that our method can significantly
improve the image quality. Ablation studies verify the effectiveness of the components in our proposed framework. In the future, we will explore more inner-structure in the sinogram for further improvement on the LDCT.

REFERENCES

[1] M. Balda, J. Hornegger, and B. Heismann, “Ray contribution masks for structure adaptive sinogram filtering,” IEEE Trans. Med. Imag., vol. 31, no. 9, pp. 1228–1239, Jun. 2011.

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

[3] G. Chen et al., “ArtNet: Fused analytical and iterative reconstruction with deep neural network regularization for sparse-data CT,” Med. Phys., vol. 47, no. 7, pp. 2916–2930, Jul. 2020.

[4] Z. Li, T. Zhang, and G. Wang, “A limited-angle CT reconstruction method based on anisotropic tv minimization,” Phys. Med. Biol., vol. 58, no. 7, p. 2119, 2013.

[5] Y. Zhang, Y. Wang, W. Zhang, F. Lin, Y. Pu, and J. Zhou, “Statistical iterative reconstruction using adaptive fractional order regularization,” Biomed. Opt. Exp., vol. 7, no. 3, pp. 1015–1029, 2016.

[6] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, “Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation,” in Proc. 33rd AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 865–872.

[7] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, “Auto-GAN: Self-supervised collaborative learning for medical image synthesis,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 10486–10493.

[8] S. Miao et al., “Dilated FCN for multi-agent 2D/3D medical image registration,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 4694–4701.

[9] P. Fan, F. Chen, C. Liu, W. Kong, and D. Zhang, “Hierarchical temporal attention network for thyroid nodule recognition using dynamic CEUS imaging,” IEEE Trans. Med. Imag., vol. 40, no. 6, pp. 1646–1660, Jun. 2021.

[10] Z. Chen, T. Zhang, P. Wan, and D. Zhang, “SEGAN: Structure-enhanced generative adversarial network for compressed sensing MRI reconstruction,” in Proc. AAAI Conf. Artif. Intell., vol. 33, Jul. 2019, pp. 1012–1019.

[11] Y. Chen et al., “Improving low-dose abdominal CT images by weighted intensity averaging over large-scale neighborhoods,” Eur. J. Radiol., vol. 89, no. 9, pp. e42–e49, Nov. 2011.

[12] Y. Chen et al., “Thoracic low-dose CT image processing using an artifact suppressed large-scale nonlocal means,” Phys. Med. Biol., vol. 57, no. 9, p. 2667, 2012.

[13] S. Ha and K. Mueller, “Low dose CT image restoration using a database of image patches,” Phys. Med. Biol., vol. 60, no. 2, p. 869, 2015.

[14] Z. Zhang, X. Han, E. Pearson, C. Pelizzari, E. Y. Sidky, and X. Pan, “Artifact reduction in short-scan CBCT by use of optimization-based reconstruction,” Phys. Med. Biol., vol. 61, no. 9, p. 3387, 2016.

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

[16] H. Shan et al., “Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction,” Nature Mach. Intell., vol. 1, no. 6, pp. 269–276, 2019.

[17] X. Yin et al., “Domain progressive 3D residual convolution network to improve low-dose CT imaging,” IEEE Trans. Med. Imag., vol. 38, no. 12, pp. 2903–2913, Dec. 2019.

[18] Y. Zhang et al., “CLEAR: Comprehensive learning enabled adversarial reconstruction for suble structure enhanced low-dose CT imaging,” IEEE Trans. Med. Imag., vol. 40, no. 11, pp. 3089–3101, Nov. 2021.

[19] Q. Xie et al., “Robust low-dose CT sinogram preprocessing via exploiting noise-generating mechanism,” IEEE Trans. Med. Imag., vol. 36, no. 12, pp. 2487–2498, Dec. 2017.

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

[21] P. T. Lauzier and G.-H. Chen, “Characterization of statistical prior image constrained compressed sensing (PICCS): II. Application to dose reduction,” Med. Phys., vol. 40, no. 2, Jan. 2013, Art. no. 021902.

[22] X. Zheng, S. Ravishankar, Y. Long, and J. A. Fessler, “PWLS-ULTRA: An efficient clustering and learning-based approach for low-dose 3D CT image reconstruction,” IEEE Trans. Med. Imag., vol. 37, no. 6, pp. 1498–1510, Jun. 2018.

[23] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 184–199.

[24] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Esgum, “Generative adversarial networks for noise reduction in low-dose CT,” IEEE Trans. Med. Imag., vol. 36, no. 12, pp. 2536–2545, Dec. 2017.

[25] D. Hu et al., “Hybrid-domain neural network processing for sparse-view CT reconstruction,” IEEE Trans. Radiat. Plasma Med. Sci., vol. 5, no. 1, pp. 88–98, Jan. 2021.

[26] W. Wang et al., “An end-to-end deep network for reconstructing CT images directly from sparse sinograms,” IEEE Trans. Comput. Imag., vol. 6, pp. 1548–1560, 2020.

[27] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” OpenAI, San Francisco, CA, USA, Tech. Rep., 2018. [Online]. Available: https://www.cs.stanford.edu/~amfradford/language-improving.pdf

[29] A. Radford et al., “Language models are unsupervised multitask learners,” OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[31] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

[32] B. Wu et al., “Visual transformers: Token-based image representation and processing for computer vision,” 2020, arXiv:2006.03677.

[33] A. Hassan, S. Walton, N. Shah, A. Abuداءwi, J. Li, and H. Shi, “Escaping the big data paradigm with compact transformers,” 2021, arXiv:2104.05704.

[34] H. Chen et al., “Pre-trained image processing transformer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12299–12310.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 213–229.

[36] S. Boyd, S. P. Boyd, and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[37] T. Sun, N. Sun, J. Wang, and S. Tan, “Iterative CBCT reconstruction using Hessian penalty,” Phys. Med. Biol., vol. 60, no. 5, p. 1965, 2015.

[38] L. Yu, M. Shiung, D. Dondal, and C. H. McCollough, “Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols,” in Proc. Comp. Tomog. Med., vol. 36, no. 4, pp. 477–487, 2012.

[39] R. Anirudh, K. Kim, J. J. Thiagarajan, K. A. Mohan, K. Champley, and T. Bremer, “Lose the views: Limited angle CT reconstruction via implicit sinogram completion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 902–911.

[40] W. van Aarle et al., “Fast and flexible X-ray tomography using the ASTRA toolbox,” Opt. Exp., vol. 24, no. 22, pp. 25129–25147, 2016.

[41] K. L. Boedecker and M. F. McNitt-Gray, “Application of the noise power spectrum in modern diagnostic MDCT: Part I. Measurement of noise power spectrum and noise equivalent quanta,” Phys. Med. Biol., vol. 52, no. 14, p. 4027, 2007.