Journal of Digital Imaging (2023) 36:2290–2305
https://doi.org/10.1007/s10278-023-00842-9

HCformer: Hybrid CNN-Transformer for LDCT Image Denoising

Jinli Yuan[1] · Feng Zhou[1] · Zhitao Guo[1] · Xiaozeng Li[1] · Hengyong Yu[2]

Received: 24 January 2023 / Revised: 29 April 2023 / Accepted: 2 May 2023 / Published online: 29 June 2023

Abstract

Low-dose computed tomography (LDCT) is an effective way to reduce radiation exposure for patients. However, it will increase the noise of reconstructed CT images and affect the precision of clinical diagnosis. The majority of the current deep learning-based denoising methods are built on convolutional neural networks (CNNs), which concentrate on local information and have little capacity for multiple structures modeling. Transformer structures are capable of computing each pixel's response on a global scale, but their extensive computation requirements prevent them from being widely used in medical image processing. To reduce the impact of LDCT scans on patients, this paper aims to develop an image post-processing method by combining CNN and Transformer structures. This method can obtain a high-quality images from LDCT. A hybrid CNN-Transformer (HCformer) codec network model is proposed for LDCT image denoising. A neighborhood feature enhancement (NEF) module is designed to introduce the local information into the Transformer's operation, and the representation of adjacent pixel information in the LDCT image denoising task is increased. The shifting window method is utilized to lower the computational complexity of the network model and overcome the problems that come with computing the MSA (Multi-head self-attention) process in a fixed window. Meanwhile, W/SW-MSA (Windows/Shifted window Multi-head self-attention) is alternately used in two layers of the Transformer to gain the information interaction between various Transformer layers. This approach can successfully decrease the Transformer's overall computational cost. The AAPM 2016 LDCT grand challenge dataset is employed for ablation and comparison experiments to demonstrate the viability of the proposed LDCT denoising method. Per the experimental findings, HCformer can increase the image quality metrics SSIM, HuRMSE and FSIM from 0.8017, 34.1898, and 0.6885 to 0.8507, 17.7213, and 0.7247, respectively. Additionally, the proposed HCformer algorithm will preserves image details while it reduces noise. In this paper, an HCformer structure is proposed based on deep learning and evaluated by using the AAPM LDCT dataset. Both the qualitative and quantitative comparison results confirm that the proposed HCformer outperforms other methods. The contribution of each component of the HCformer is also confirmed by the ablation experiments. HCformer can combine the advantages of CNN and Transformer, and it has great potential for LDCT image denoising and other tasks.

Keywords Low-dose CT · Deep learning · CT image denoising

Introduction

Due to the ability that Computed Tomography (CT) can rapidly obtain a high-resolution images to capture structural features and pathological conditions [1], it is widely used in medical diagnosis, image-guided surgery, and radiation therapy [2–4]. However, the radiation generated during CT scanning is harmful to the human body and increases the risk of cancers. Research on low-dose computed tomography (LDCT) has attracted considerable attention [5]. Radiation

✉ Zhitao Guo
mrnow@hebut.edu.cn

Jinli Yuan
jinli_yuan@hebut.edu.cn

Feng Zhou
1226406144@qq.com

Xiaozeng Li
33574624@qq.com

Hengyong Yu
hengyong_yu@uml.edu

[1] The School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China

[2] The Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854, USA

dose reduction can be achieved by reducing the number of projection attempts and/or lowering the X-ray flux, but the reduction of radiation dose may lead to poorer image quality, which is usually manifested as noises and artifacts in the resulting images [6].

There are mainly three types of methods available to improve the quality of LDCT images: projection domain filtering algorithms, iterative reconstruction algorithms, and post-processing methods [7]. Among those three types of methods, a more direct approach is to use image post-processing to suppress noise in LDCT images. However, it is still challenging due to the ill-posed nature of the noise. Compared with the traditional LDCT image denoising methods, the convolutional neural network (CNN) has achieved better performance in learning generalized prior. Chen et al. [8] first applied CNNs to the post-processing of LDCT, using a deep CNN structure to map LDCT images to the corresponding Normal-Dose CT (NDCT) images. It showed certain advantages in terms of visual assessment and quantitative metrics compared to conventional methods. Chen et al. [9] proposed a RED-CNN network with residual coding and decoding structure for better denoising results. Wu et al. [10] proposed a cascaded CNN denoising algorithm, which can obtain a higher quality denoised images compared with conventional CNN algorithms. Wolterink et al. [11] applied the Generative Adversarial Network (GAN) to LDCT image denoising. Yang et al. [12] used the Wasserstein distance in GAN to form a WGAN network for LDCT denoising. Yi and Babyn [13] used a sharpness detection network to guide the training process of the network and obtained a small resolution loss, which improved the LDCT image quality. Yin et al. [14] proposed the DP-ResNet for LDCT denoising based on an asymptotic 3D residual convolutional network. Guo et al. [15] combined edge gradient information with a multi-stage network for LDCT image denoising and achieved good results.

The aforementioned CNN-based works have obtained inspiring results by refining different aspects of the network structures. For CNN Networks, feature learning is based on convolution operation, which brings local connectivity and strong feature learning capability. Although these capacities provide CNNs with efficiency and versatility, there are two main problems. First, the convolution operation is constrained by a limited field of perception, resulting in too much focus on local features. This disadvantage makes it difficult for CNN-based methods to make full use of the similarity between the large regions in the image. Hence, CNN-based methods are inefficient to model various information in CT images [32]. Second, the convolution kernels have static weights during the operations, and using the same convolution kernel to recover different parts of the image may lose details and result in a poorly denoised image. Because of the particularity of LDCT images, it is necessary to consider the complete characteristics of different tissues in the images, which puts forward the higher requirements for the long-range correlations of images.

In recent years, Transformer structure with Multi-head self-attention (MSA) as the core mechanism has been successfully applied [16]. Weighted by other locations to compute the response for a given pixel location [17, 18], the MSA mechanism can dynamically calculate the correlation of pixels at all locations in an image. This can provide a better alternative for LDCT image denoising tasks [19]. Transformer has achieved an excellent performance in image processing fields such as image classification [20, 21], image segmentation [22], and object detection [23, 24]. The MSA mechanism in Transformer enables content-based interaction between image content and attention weights. This process can be interpreted as achieving spatially varying convolution [25–27]. It makes the Transformer very effective in capturing the features in different structures, particularly for image denoising tasks. Chen et al. [28] proposed a pre-trained image denoising model based on Transformer (IPT), and for the first time applied it to the field of image denoising. However, due to the computational limitations of Transformer, IPT adopts a method to divide the input image into 48×48 small patches. Too small input patches may lead to information loss and the generation of boundary artifacts [29]. In the field of medical image processing, Transformer has been successfully applied in some tasks, such as medical image segmentation and nodule detection [30, 31]. LDCT image denoising requires to denoise different information between various organizational structures in the images. Transformer has also been successfully used in this field. For example, Zhang et al. [32] proposed TransCT by using different characteristics of a high frequency and low frequency in LDCT images, and Qiao et al. [33] proposed a CTC method by combining Transformer with edge enhancement. Luthra et al. [34]. proposed an edge enhancement Transformer model for medical image denoising, which combines the learnable Sobel filter for edge enhancement to improve the performance of the overall architecture.

The existing problems can be summarized as follows: (1) because CNN is limited by the perceptual field and the static weights at the time of operation, it cannot effectively use the similarity among large regions in LDCT images. This will lead to the destruction of image details after the LDCT image denoising, and the information between tissues in LDCT images cannot be effectively recovered. (2) Because the standard Transformer structure has high computational complexity, it cannot be effectively applied for high-resolution LDCT image denoising. Meanwhile, the Transformer structure mainly focuses on extracting global spatial information between pixels, ignoring the importance of pixel neighborhood information for image denoising. (3) The mean square error (MSE loss) of individual

pairs is often used as a conventional objective function in many LDCT denoising algorithms to achieve higher signal-to-noise ratios. However, this pixel-by-pixel comparison method tends to cause a loss of detailed information and thus over-smoothing the recovery results.

For the above problems, we propose a hybrid CNN-Transformer codec network, named HCformer. This framework can process the features of the input image with full resolution, avoiding information loss that may be caused by up- and down-sampling in traditional U-shaped structures. For LDCT image denoising tasks, pixel neighborhood information plays an important role [35, 36]. And in medical image processing, the expression of neighborhood information is also critical. The neighborhood feature enhancement (NFE) module proposed in this paper uses a hybrid structure to replace the multiple-layer perceptron (MLP) layers in the standard Transformer layer. Specifically, by refining the convolution in the depth direction and spatial dimension of the NEF module, the local information extracted by the convolution operation can be introduced into the Transformer process. The hybrid CNN-Transformer structure ensures that the Transformer is complemented by local information in different channels and spatial dimensions when modeling the global information relevance of the image.

Furthermore, to reduce the computational burden, we use the window/sliding-window MSA (W/SW-MSA) mechanism to connect two consecutive self-attention layers by sliding window, which was proved to be more effective than using a simple sliding window [37]. By using two consecutive layers of different windows, the information interaction between the overall windows of the image can be obtained efficiently, and this can avoid the information wastage that may result from stationary windows. Using the hybrid codec structure of CNN and Transformer with the W/SW-MSA, the computational burden can be reduced. The respective advantages of CNN and Transformer in extracting features can be effectively exploited. These methods effectively enable the denoising of a high-resolution LDCT images.

Our main contributions are threefold. (1) This paper proposes a hybrid codec network based on CNN-Transformer, which adopts Transformer as the main structure to realize the mapping from LDCT to NDCT images. Meanwhile, the hybrid structure is also reflected in the Transformer layer operation level. The NFE module with multidimensional CNN is proposed to replace the MLP layer in the standard Transformer layer. Hence, the local information can be introduced into global-level operation, making the network more suitable for LDCT image denoising. (2) Transformer is used as the main structure to implement the mapping from LDCT to NDCT images. To avoid excessive computational burden caused by the traditional Transformer during computation, during the construction of Transformer codec structure in HCformer, the information exchange between different windows is realized by the W/SW-MSA mechanism. This method can effectively reduce the computation of the Transformer and makes the network more flexible for LDCT denoising. (3) The loss of detailed information caused by the single MSE loss pixel-by-pixel comparison method may blur the recovery results. A loss function, including Edge loss and MSE loss, is used as the final loss function to guide the network training.

Method

Denoising Model

The LDCT images have more noise and poor image quality. In this paper, the deep learning method is used to remove noise for high-quality CT images. The problem of LDCT image denoising can be mathematically modeled as follows. Assuming $X \in R^{m \times n}$ is the LDCT image and $Y \in R^{m \times n}$ is the corresponding NDCT image, we have

$$X = \sigma(Y), \tag{1}$$

where σ represents the degrading process with noise. The denoising from LDCT image X to NDCT image Y in Eq. (1) can be treated as an inverse mapping of σ. The goal is to find a function F that maps LDCT image X to NDCT image Y, to minimize the cost function:

$$\underset{F}{\operatorname{argmin}} ||F(X) - Y||_2^2, \tag{2}$$

where F is the best approximation of σ^{-1} and represents a deep learning neural network based on a learnable Transformer. The detailed process is described in the next subsection.

Overall Network Structure

The goal of this paper is to design an efficient CNN-Transformer hybrid network for LDCT image denoising, which can effectively learn the features of a high-resolution LDCT images and implement the mapping from LDCT to NDCT images. The convolutional layer has a satisfactory effect in early visual processing, which is attractive to more stable optimization and better results [38]. The Transformer is excellent at extracting global information, and as the main decoder-encoder structure, it can effectively make up for the problem that the convolution operation has insufficient computing capability for remote pixel correlation. HCformer is a hybrid codec structure composed of CNN and Transformer. It mainly consists of four parts, including a feature extraction block (FEB), encoder module, decoder module, and image restoration module (IRB), as shown in Fig. 1.

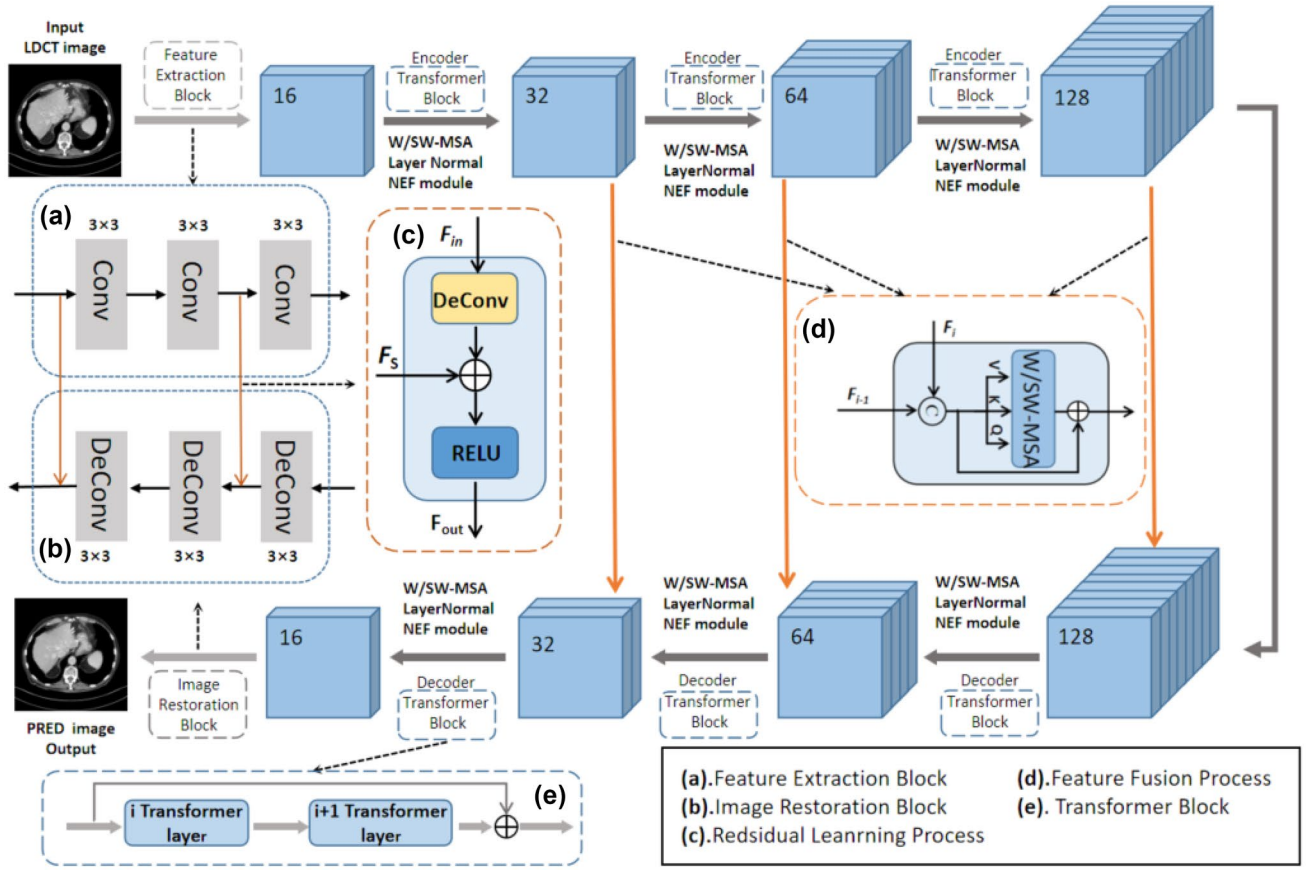

Fig. 1 The overall network structure

The feature extraction block can extract the input image features step by step through three layers of convolution and use them as the input of the encoder, as shown in Fig. 1a. The encoder and decoder modules use symmetric Transformer Blocks to implement the encoding and decoding process. The W/SW-MSA mechanism is used to reduce computational costs during codec mapping. Meanwhile, the NFE module is used to enhance the expression ability of local information in the network structure of each Transformer layer. The final IRB is symmetric with the feature extraction Block, and three layers of DeConv operation are used to achieve mapping from the feature domain to the image domain (see Fig. 1b).

Define the input LDCT image as $X_i \in R^{M\times N\times Cin}$ where M, N, and C_{in} are the width, height, and number of input channels, respectively. Then, after passing through the feature extraction module, it can generate the shallow feature X_F satisfying

$$X_F + X_S = F_{FE}(X_i), \tag{3}$$

where $X_F \in R^{M\times N\times C}$ is the multidimensional feature through the feature extraction module, $F_{FE}(\cdot)$ denotes the feature extraction process, C denotes the number of feature channels after the shallow feature extraction module, and X_S denotes the shortcut generated by the feature extraction. The convolution layer can map the input image into a high-dimensional space. Next, through the Transformer-based encoding structure, X_F is used as the input to the encoder, and the process is shown in Eq. (4). It is set to 16 in the initial encoding block in this paper, and then multiplied successively to finally generate 128 channel features. In the decoding process, the channels change is symmetrical with the encoding process, and the feature is finally denoised to the initial channel number.

$$X_{ES} + X_{EC} = F_{En}(X_F), \tag{4}$$

where $F_{EN}(\cdot)$ is a Transformer-based encoder structure containing i layers of Transformer blocks with feature encoding capabilities, and more specifically, intermediate features $X_1, X_2 \ldots X_{i-1}$ and the residual outputs $X_{s1}, X_{s2} \ldots X_{Si-1}$, and the output depth encoding features X_{EC} and the intermediate layer features X_{ES} required for feature fusion are extracted step-wise in the following manner:

$$X_i + X_{si} = T_{ECTB}(X_{i-1})$$

$$X_{ES} + X_{EC} = T_{ECTB}(X_i) \tag{5}$$

where $T_{ECTB}(\cdot)$ denotes the i_{th} Transformer encoding block. Using multidimensional convolution operations in the Transformer layer of the encoding block can introduce local inductive bias into the Transformer-based network. This method can lay a better foundation for fusing partial and global features.

The decoder structure is also based on a hybrid structure of the Transformer and CNN. To enhance the encoder output, this paper replaces one of the convolutional layers in the multidimensional encoding block with a deconvolution as the output of the final Transformer block, forming a symmetric operation with the convolutional layer in the encoder. The process can be described as follows:

$$X_{DO} = T_{DETB}(X_{ES} + X_{EC}) \tag{6}$$

where X_{DO} denotes the output feature of the decoder and $T_{DETB}(\cdot)$ denotes the Transformer decoding block. The output features $X_{DO} \in R^{M\times N\times C}$ after codec restoration are finally recovered by the image restoration module into a recovered image $X_{Predict} \in R^{M\times N\times Cin}$ close to the NDCT image, and the final process can be described as:

$$X_{\mathrm{Predict}} = F_{IRB}(X_{DO} + X_S) \tag{7}$$

In Eq. (7), $T_{IRB}(\cdot)$ is the restoration module, and the features initially extracted by convolution mainly contain local features, while the deep features extracted by the Transformer block focus on recovering the global features of LDCT images. Through the residual connection, HCformer can transfer the local information directly to the codec module. This helps the deep feature extraction module to focus on the global information and stabilize the training process. For the implementation of the restoration module, this paper uses three layers of deconvolution layers to gradually reduce the recovered features from the feature dimension to the image domain.

When the network structure is too deep, the feature fusion interaction strategy can avoid the interference caused by the feature fusion and effectively integrate the features of different stages and levels. At the beginning of the three-layer convolutions, the initial early features are mapped to the denoised image by adding each pixel to enhance the detail information and suppress the noise, making the ratio of different abstract feature information more suitable for the LDCT denoising and realizing more flexible feature utilization, as shown in Fig. 1b.

To effectively transfer shallow features from the encoding block to the decoding block, the feature fusion process is adopted to transfer features of different stages between the encoder and decoder based on Transformer (see Fig. 1d). The input feature F_i of the first layer is spliced together with the output F_{i-1} of the Transformer Block of the previous layer. Then, the blending feature is input before the W/SW-MSA calculates the projection of K and V.

Feature fusion interaction strategy can consider both shallow information and details. It is helpful to consider detail units in LDCT images. First, LDCT image denoising requires more analysis of small areas to underline features, which may lead to too small weight values that cannot be effectively expressed in the process of step-by-step transmission. Second, when the image needs global modeling, more attention is paid to broader information in the image, so the long-distance pixel response to the image is more dependent on the correlation between the advanced features extracted after a certain deep network.

Transformer Module Based on Encoder-decoder Structure

Transformer Layer

Figure 2 shows the Transformer layer in the Transformer Block, which is implemented by two alternating layers between window attention (W-MSA) and shifted window attention (SW-MSA), respectively. The standard Transformer structure and its adaptation to image classification are subject to a global self-attention computation, which calculates the relationship between a pixel at that location and all other pixels at that location in the input image. The

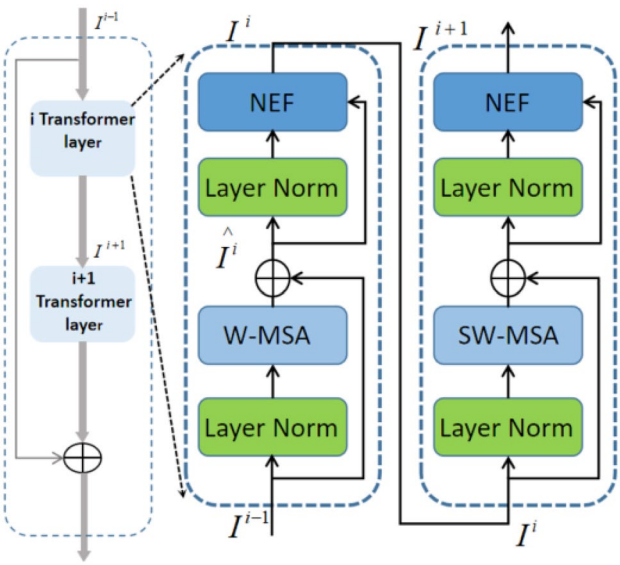

Fig. 2 Transformer layer

global computation leads to a quadratic relationship between the computational complexity and the input image size. To improve the efficiency of the model computation, this paper uses a self-attention computation by dividing non-overlapping windows. The input is adapted from the token to the feature map extracted by the feature extraction block. The size of the feature map input by the Feature Extraction Block is $H \times W$.

The computational complexity of the global MSA mechanism is a quadratic product of the input feature map sizes H and W. To perform the MSA efficiently, the W-MSA is computed inside a predefined window, which is set to segment the image in a non-overlapping manner. M is the size of each window and the number of windows is$N = HW/M^2$. The W-MSA can effectively transfer the quadratic relationship to the parameter M. The computational intensity of a single parameter M is significantly lower than that of the squared relationship of the quadratic product of H, W. The main difference in computational complexity between MSA and W-MSA can be approximated as:

$$\frac{\Omega(MSA)}{\Omega(W-MSA)} \approx \frac{2(HW)^2C^2}{2M^2HWC^2} \tag{8}$$

and it can be simplified as:

$$\frac{\Omega(MSA)}{\Omega(W-MSA)} \approx \frac{HW}{M^2} \tag{9}$$

For example, if the input feature map size is set as 128×128 and the window size is set to 8, the computational complexity of W-MSA can be reduced by about 256 times compared to MSA. This operation proves to be effective in reducing the computational burden of MSA in Transformer.

In this paper, we use two different window division methods in successive Transformer layers, as shown in Fig. 3.

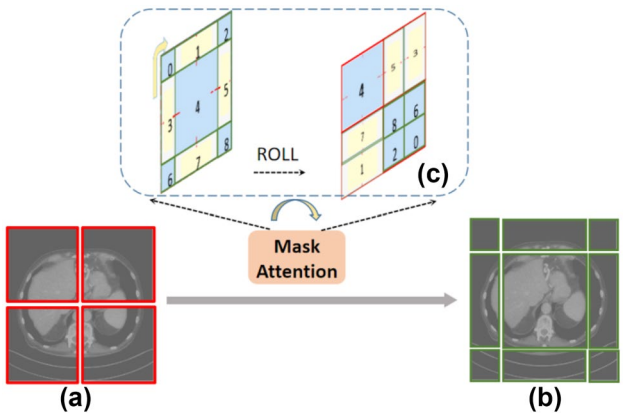

Fig. 3 Schematic diagram of the window division between layer i and layer $i+1$. **a** Transformer layer window division for layer i. **b** Transformer layer window division for layer $i+1$. **c** Process of window displacement

Relying only on the non-overlapping and fixed local window self-attention mechanism, it will result in no information exchange between different windows. To maintain efficient computation of non-overlapping windows and introduce the cross-window information supplementation, the shifted-window method is used to implement the window division from a to b in Fig. 3. Since the window division in Fig. 3b is non-regular, it is too complicated to simply rotate and move the window. In this paper, the feature map is shifted by 1/2 window unit in the direction of the center point and then overlapped with a fixed serial port to obtain the window in b. This operation also relies on a cyclic displacement operation to achieve the filling of the feature area in Fig. 3c with Mask on the corresponding position.

To avoid the information interference of unconnected regions in the W-MSA process, the Mask-Attention is used to calculate the correlations in the same regions. Specifically, in the process of computing Attention, let Q and K be computed with the same label. Then, this operation ignores the correlation between unrelated indices, to avoid the weights brought by non-related regions. Benefiting from the broadcast operation in PyTorch, we can easily get the mask feature map of the same dimension, and the weight value is retained according to the need when the dimension is obtained with the same feature region. The attention mask and the discriminant function output are used in different windows to separate the information between different windows and prevent the information confusion that may be caused when calculating MSA during the loop. Mask-Attention is used to solve the problem of irregular window division, and the Attention calculation process in Fig. 3 is equivalent to realizing the information interaction among different Windows.

NFE Module

In the process of LDCT image denoising, adjacent pixels also play an important role in image denoising. To enhance the expression of local association information in the denoised image, a neighborhood feature enhancement module (NFE module) is proposed in the Transformer layer to replace the MLP in the standard Transformer layer, as shown in Fig. 4. Since the image recovery task focuses more on the relationship between input pixels and mapped pixels, the corresponding standard MLP layer focuses more on the output discriminative information. We use the convolution operation in the NFE module to enhance the fine-grained nature of the recovered image.

The NFE module uses the Depth-convolution of channel direction to provide channel-level pixel aggregation of local context information, introducing local information as a supplement to the Transformer layer attention operation. After

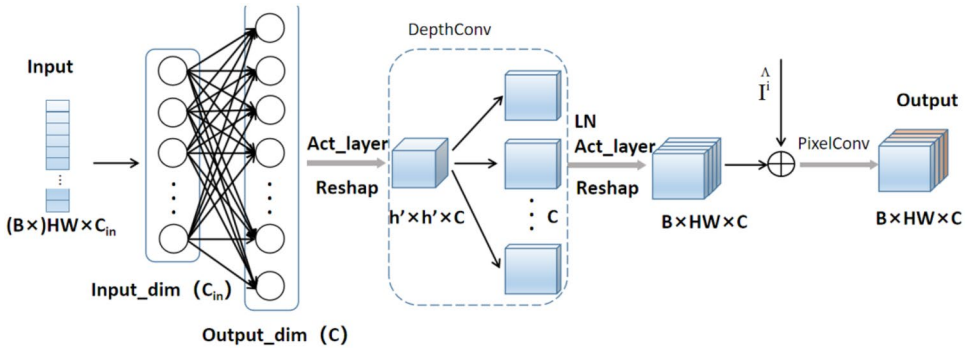

Fig. 4 Neighborhood feature enhanced module

the introduction of MSA global weights, the cross-channel pixel-level information aggregation of output features is carried out by using super pixel-level convolution. Pixel convolution is slightly different in encoding and decoding. In decoding, 1×1 convolution is used to refine the feature map pixel by pixel. The same DeConv was used in the decoding block to symmetrically decode the feature pixel by pixel. This operation ensures that the Transformer models the global information relevance of the image and complements it with different levels of local information.

The input features are first mapped into the hidden channel by a fully connected layer (output dim in Fig. 4), which is expressed as C in the subsequent operation. The input spreading features are then expanded into a feature map of shape $$m\in {{{R}^{h}}^{'}}^{\times h'\times C}$$ by an activation function as well as reshaping, where $$h\text{'}=sqrt\left(HW\right)$$. Next, a 3×3 depth convolution is used to capture the local information, and the convolution is performed in the channel depth direction based on C channels. This operation allows the correlation of the image channel depth to be extracted while reducing the number of parameters based on channel grouping. After obtaining the local features by deep convolution, the features are recovered to $$h\text{'}=sqrt\left(HW\right)$$ dimensions by linear transformation layer and recovery shaping to the extending features. Then, the attention weights after multi-head attention are added to the output pixel by pixel to obtain the output that focuses on both global high-dimensional features and local low-order information.

The introduction of the NEF module into the self-attention operation process can make full use of local image information. The i^{th} layer Transformer Layer and $(i+1)^{th}$ layer operation process are in Eq. (10).

$$\begin{aligned}
&F_{map} = \left\{F_{\mathrm{map}}{}^{1}F_{\mathrm{map}}{}^{2},\ldots,F_{\mathrm{map}}{}^{i}\right\},\mathrm{i} = \frac{HW}{M^2}\\
&F^{\mathrm{i}}_{W-MSA} = W-SMA\left(F^{\mathrm{i}-1}_{map}W^{Q}_{k},F^{\mathrm{i}-1}_{map}W^{K}_{k},F^{\mathrm{i}-1}_{map}W^{V}_{k}\right)+F^{\mathrm{i}-1}_{map}\\
&F^{\mathrm{i}}_{NFE} = NFE\left(FC\left(F^{\mathrm{i}}_{W-MSA}\right)\right)\\
&F^{\mathrm{i}+1}_{SW-MSA} = SW-MSA\left(F^{\mathrm{i}}_{NFE}W^{Q}_{k},F^{\mathrm{i}}_{NFE}W^{K}_{k},F^{\mathrm{i}}_{NFE}W^{V}_{k}\right)+F^{\mathrm{i}}_{NFE}\\
&F^{\mathrm{i}+1}_{NFE} = NFE\left(FC\left(F^{\mathrm{i}+1}_{SW-MSA}\right)\right)
\end{aligned} \tag{10}$$

where F_{map} denotes the Feature map input to the i^{th} Transformer layer, assuming that the head number is k^{th} and the dimension corresponding to each head is $d_k = C/k$, where W^{Q}_{k}, W^{K}_{k}, $W^{V}{}_{k} \in R^{C\times d_k}$ denote the projection matrix of query, key, and value of the k^{th} head, respectively.

Loss Function

In this paper, the Loss function adopts a composite form. The first part is a mean square error (MSE)

$$L_{MSE}(X_S,Y) = \sqrt{||X_S - Y||^2} \tag{11}$$

As the traditional objective function in neural network-based image-denoising tasks, MSE is often used in many LDCT image-denoising algorithms to achieve a higher signal-to-noise ratio. However, this pixel-by-pixel comparison approach is prone to loss of detailed information, resulting in blurred results. To avoid over smoothed images caused by a single loss function, we also introduce an Edge loss function L_{edge}

$$L_{edge}(X_s,Y) = \sqrt{||\Delta(X_s) - \Delta(Y)|| + \varepsilon^2} \tag{12}$$

where Δ represents the Laplace operator. The constant ε is to avoid singularity, and it is empirically set to 10^3. Hence, the composite loss function can be defined as:

$$L_{Compound} = L_{MSE}(X_s,Y) + \lambda L_{edge}(X_s,Y) \tag{13}$$

where λ is to balance the two terms.

By introducing the Laplace operator into the loss function, the two-dimensional isotropy of the second spatial derivative of the image can be introduced into the training process of the supervised network. Using Laplace operation can improve the denoising ability of the network for the abrupt position of response intensity in the image. These positions are usually represented by the boundaries of different tissues in LDCT images and parts with complex and varied structures. The denoising and

Table 1 Quantitative comparison of different network structures

	SSIM (Mean ± STD)	HuRMSE (Mean ± STD)	FSIM (Mean ± STD)
LDCT	0.8017 ± 0.0453	34.1898 ± 8.7348	0.6885 ± 0.0284
Full-CNN	0.8275 ± 0.0406	19.1232 ± 4.4397	0.7042 ± 0.0260
With MLP	0.8324 ± 0.4288	18.0542 ± 4.3232	0.7177 ± 0.0259
HCformer(ours)	**0.8507 ± 0.0405**	**17.7213 ± 4.3480**	**0.7247 ± 0.0255**

The bold font indicate that these results are optimal

reconstruction effects of these key positions are crucial for the denoising task of LDCT images. As shown in Eq. (13), the final loss function is a weighted combination of MSE loss and Edge loss. The combined loss function is used to guide the network learning, which enables the network to learn high-order information implicit in the images and the details of the edge information with a similar degree between the predicted and prior images.

Experiments and Results

Dataset

A clinical dataset from the 2016 NIH AAPM Mayo Clinic Low-Dose CT Grand Challenge, licensed by Mayo Clinics, is used to train and test the network. The dataset consists of projection data and reconstructed images with a slice thickness of 3 mm. There are 2378 normal dose CT image slices and their corresponding low-dose (quarter-dose) CT image slices from 10 anonymous patients, each of which includes 512×512 pixels.

In the training process, a data expansion technology is used to extract 128×128 image patches from the random position and random sequence of each slice. Random rotation and flipping operations are also used for these image blocks to improve the diversity of input samples and reduce the correlation between training samples. For a given patient case, although the noises in the same position in all the image slices might have correlation, it should be a different story for different patches at different positons. For testing, the trained model is directly applied to full image slices of the test patient scans. The network proposed in this paper is a fully convolutional network, which can use image patches for training and testing full-size images (512×512). Therefore, the image patches strategy can help to reduce the amount of computational cost, improve the training speed, and increase the number of training samples to avoid overfitting.

Experimental Setup

The experiments are carried out on a 64-bit Windows 10 system with PyTorch deep learning framework for network training and CUDA-Toolkit10.1 for its acceleration. The hardware configuration is the Intel Core i7-9700 K CPU@3.2 GHz, 16 G of RAM, and NVIDIA GeForce RTX series. The weights are iteratively optimized using the Adam optimization algorithm during the training process.

Measurement Metrics

To evaluate the effectiveness of the proposed algorithm, structural similarity (SSIM), Hu root mean square error (HuRMSE), and feature similarity (FSIM) are used to evaluate image quality. SSIM evaluates the structural differences between denoised and real images, and it is typically used to characterize the degree of detail similarity between images. FSIM represents the similarity of visual feature structures. HuRMSE is used as an evaluation indicator for the distance between denoised LDCT and NDCT images. The higher the values of SSIM and FSIM are, the higher the image quality is. The calculation of those metrics are as follows:

$$SSIM = \frac{(2\mu_z\mu_x + c_1)(2\delta_{z,x} + c_2)}{(\mu_z{}^2 + \mu_x{}^2 + c_1)(\delta_z{}^2 + \delta_x{}^2 + c_2)} \tag{14}$$

$$HuRMSE = \sqrt{\frac{\left(I_x - I_Y\right)^2}{N}} \tag{15}$$

$$FSIM = \frac{\sum_{x\in\Omega} S_L(x) PC_m(x)}{\sum_{x\in\Omega} PC_m(x)} \tag{16}$$

Table 2 Quantitative comparison of different input patch size

Size	SSIM (Mean ± STD)	HuRMSE (Mean ± STD)	FSIM (Mean ± STD)
64×64	0.8448 ± 0.0402	17.8578 ± 4.3726	0.7223 ± 0.0261
80×80	0.8472 ± 0.0403	17.8427 ± 4.3308	0.7237 ± 0.0253
96×96	0.8453 ± 0.4002	17.8101 ± 4.3602	0.7236 ± 0.0257
112×112	0.8449 ± 0.0403	17.8215 ± 4.3682	0.7238 ± 0.0259
128×128	**0.8507 ± 0.0405**	**17.7213 ± 4.3480**	**0.7247 ± 0.0255**

The bold font indicate that these results are optimal

Table 3 Quantitative comparison of different λ in loss function

λ	SSIM (Mean ± STD)	HuRMSE (Mean ± STD)	FSIM (Mean ± STD)
Without ELoss	0.8437 ± 0.0409	17.7327 ± 4.3476	0.7235 ± 0.0256
0.01	0.8449 ± 0.0401	17.7511 ± 4.3372	0.7247 ± 0.0252
0.02	0.8443 ± 0.0408	17.7509 ± 4.3892	0.7210 ± 0.0264
0.03	0.8459 ± 0.0403	17.7503 ± 4.3538	0.7247 ± 0.0255
0.04	0.8454 ± 0.0405	17.7412 ± 4.3489	0.7247 ± 0.0256
0.05	**0.8507 ± 0.0405**	**17.7213 ± 4.3480**	**0.7247 ± 0.0255**
0.06	0.8327 ± 0.0435	17.9795 ± 4.3386	0.7186 ± 0.0261

The bold font indicate that these results are optimal

where I_x and I_y denote the pixel Hu values of the input LDCT image and predicted LDCT image. x and z denote the LDCT image and the denoised image, and μ_x and μ_z denote the mean values of images x and$z$. δ_x and δ_z denote the standard deviation of the images x, and z and $\delta_{z,x}$ denotes the covariance between x and z. c_1 and c_2 are two default SSIM parameters defined as $(0.01 \times I_{\max})^2$ and$(0.03 \times I_{\max})^2$, where I_{max} is the maximum Hu value. z_i and x_i denote the Hu values of each pixel point of the denoised image and the LDCT image, respectively. The similarity between the predicted images and NDCT is measured using HuRMSE, and the Hu values corresponding to two CT images are directly used to determine the similarity between the LDCT and NDCT images. In Eq. (16), $PC(\cdot)$ is the phase consistency feature extracted by an orthogonal filter, and $S_L(\cdot)$ is the fusion similarity between gradient features and consistency features.

Ablation Experiments

Impact of Different Components of the Network

To investigate the role of different components in improving the quality of LDCT images, we validate several key elements of the proposed network structure and conduct experiments on the AAPM dataset. First, to verify that the introduction of the Transformer is effective, the layers in the Transformer Block of the proposed network structure are replaced with the corresponding convolutional layers, keeping the other parts unchanged. We name it as Full-CNN network. The similarity in spatial structure can also be seen by SSIM, with an improvement of about two percentage points. Meanwhile, to verify the effectiveness of the introduction of the NEF module, we also study the effect of using the MLP layer as the output of the Transformer layer and HCformer. Compared with using the MLP layer as the output, the experimental results of the NEF module have improved SSIM and HuRMSE. The results in Table 1 confirm that the proposed network can better recover LDCT images.

Impact of the Different Input Patch Size

To increase the sample size, we adopt the method of patch training and expand the sample data through inversion, rotation, and other methods. To verify the influence of different input patch sizes, we gradually increase the patch size of the Transformer from the commonly used input patch size of 64×64, and the results are summarized in Table 2.

It can be seen that the increase in patch size does not always increase the image quality, and the image quality changes are not significant. To balance the computational cost and performance, the default patch size is set as 128×128 in this paper.

The Parameterλ in the Loss Function

To verify the effectiveness of the compound Loss function (MSE Loss + Edge Loss), we provide a comparison between the effects of single MSE Loss and compound Loss. At the same time, we also evaluate the parameter λ to optimize the overall performance. As summarized in Table 3, $\lambda = 0.05$ is empirically selected based on our experiments.

Table 4 Quantitative comparison of HCformer with U-shaped codec containing upsampling and downsampling

	SSIM (Mean ± STD)	HuRMSE (Mean ± STD)	FSIM (Mean ± STD)
LDCT	0.8017 ± 0.0453	34.1898 ± 8.7348	0.6885 ± 0.0284
U-shaped codec	0.8447 ± 0.0404	17.8020 ± 4.3658	0.7234 ± 0.0258
HCformer(ours)	**0.8507 ± 0.0405**	**17.7213 ± 4.3480**	**0.7247 ± 0.0255**

The bold font indicate that these results are optimal

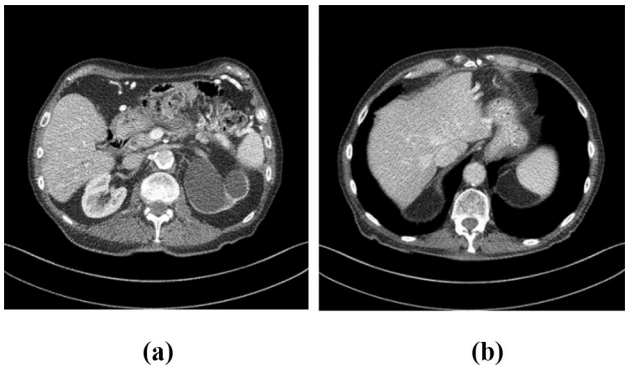

Fig. 5 Low-dose CT image samples

Comparison Between HCformer and U-Shaped Structure

In this experiment, a conventional U-shaped network structure is also constructed with the same number of layers of HCformer, containing the corresponding upsampling and downsampling operations as well as the feature fusion process. Thanks to the fact that the HCformer can effectively process full-resolution images and interact with the full-size feature maps in the network, it can be seen by the comparison in Table 4 that HCformer obtains a superior denoising effect. In general, HCformer is more suitable for LDCT image denoising, a task that focuses on pixel-by-pixel image information.

Experiment Result Analysis

Subjective Comparison

To intuitively compare the performance of the proposed algorithm, two representative sample images are selected for display. Figure 5 shows samples 1 and 2, where the observation window level is set to 40 HU and the window width is set to 400 HU.

The algorithms selected for comparison include RED-CNN [9], WGAN-VGG [12], EDCNN [39] network, MPRNet [40], and UFormer [41]. The results are evaluated using both subjective effects and objective metrics. The visual subjective effect is used to highlight the structural form of the image, and then the objective metrics are used to evaluate the degree of merit of the algorithms.

Figure 6 shows the overall comparison of Fig. 5a, and Fig. 7 shows the noise residual images after denoising by different algorithms. Figure 7 can intuitively show the difference between the denoised and NDCT images, and we can see that the denoising results based on Transformer are generally

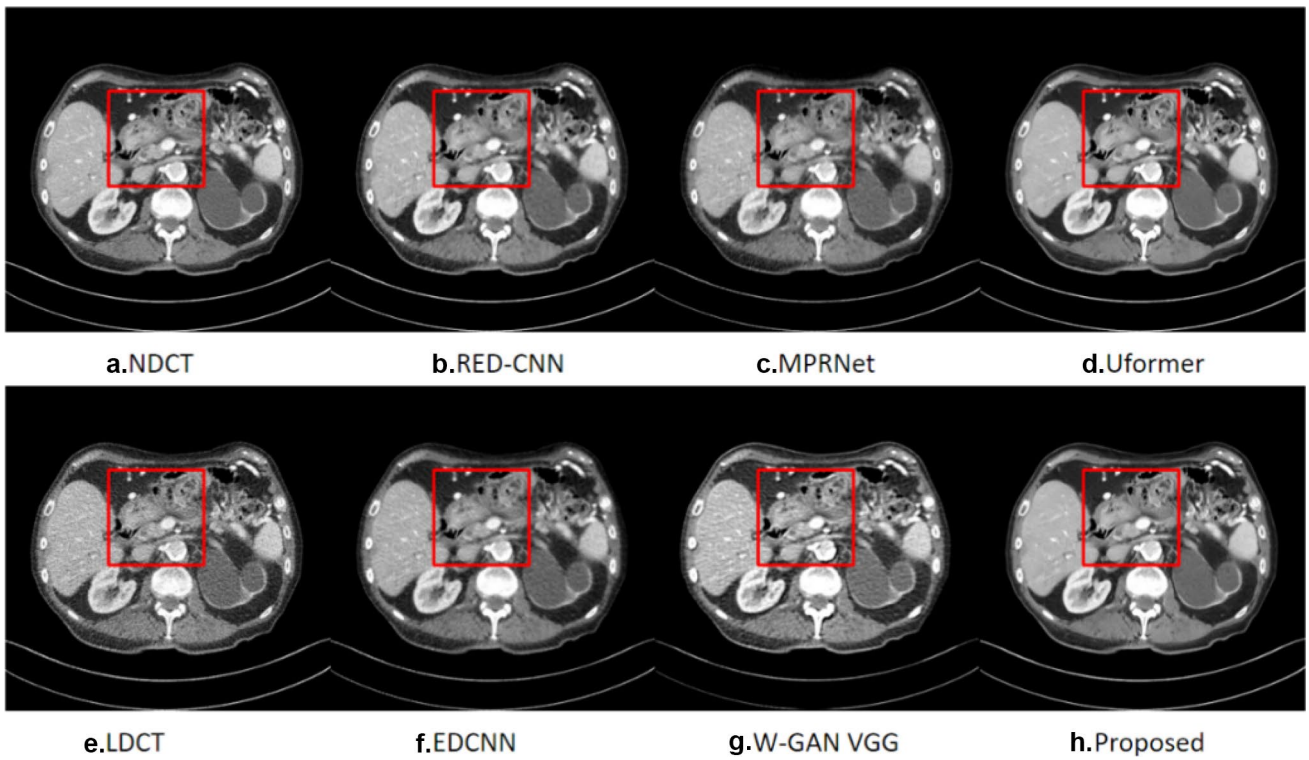

Fig. 6 Overall comparison of Sample 1. **a** NDCT image. **b** RED-CNN result. **c** MPRNet result. **d** UFormer result. **e** LDCT image. **f** EDCNN result. **g** WGAN-VGG result. **h** The result of the proposed approach

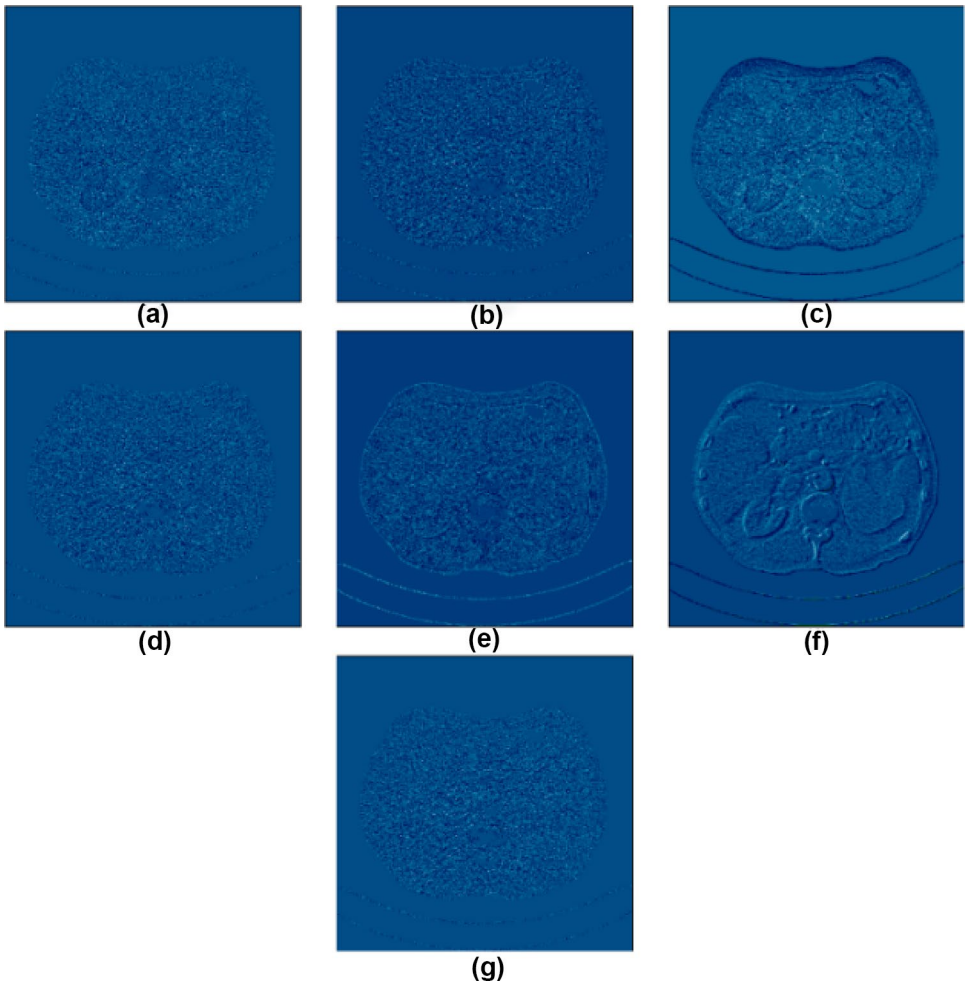

Fig. 7 Noise distribution images of sample 1 after de-noising by LDCT and different algorithms. **a** LDCT. **b** RED-CNN. **c** MPR-Net. **d** Uformer. **e** EDCNN. **f** WGAN-VGG. **g** Proposed

better than those based on CNN. The noise level in the noise image after denoising shown in Fig. 7g is lower than that of UFformer, which is the closest one to the NDCT image.

Figure 8 shows the ROI comparison for the red-box area in Fig. 6. It can be seen that all the comparison algorithms have different degrees of denoising effect. More detailed

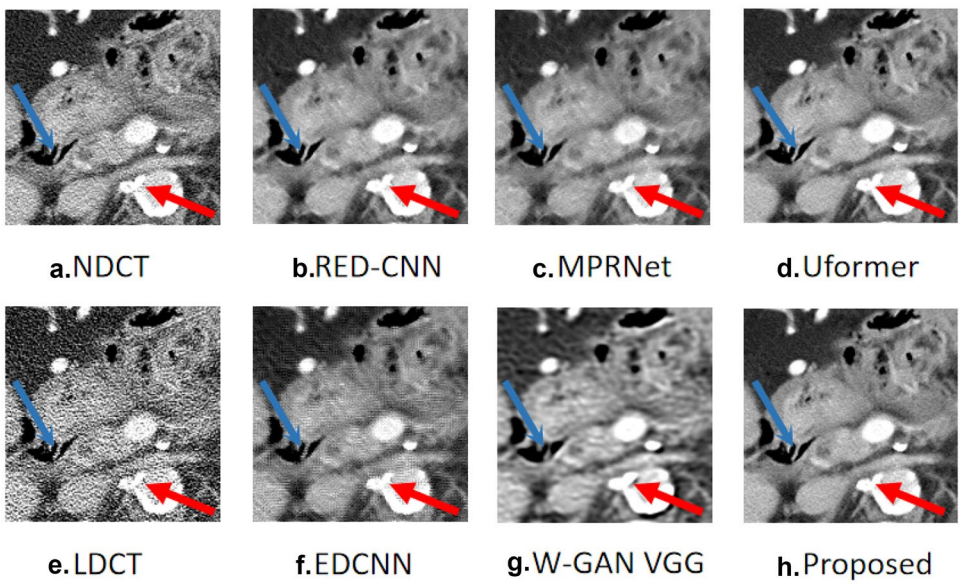

Fig. 8 The same as Fig. 6 but for magnified ROI

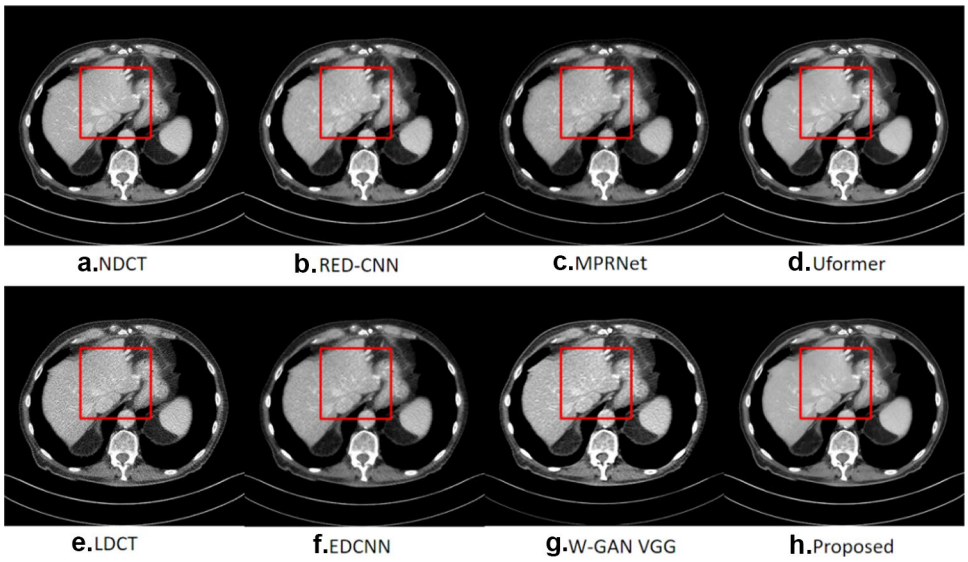

Fig. 9 Same as Fig. 6 but for sample 2

information can be found in Fig. 8. Figures 8d and h are denoised results based on the Transformer methods. Compared with other CNN-based denoising algorithms, the Transformer algorithms have a better denoising effect on LDCT images with complex structures. By comparison, it can be found that the HCformer can recover more image

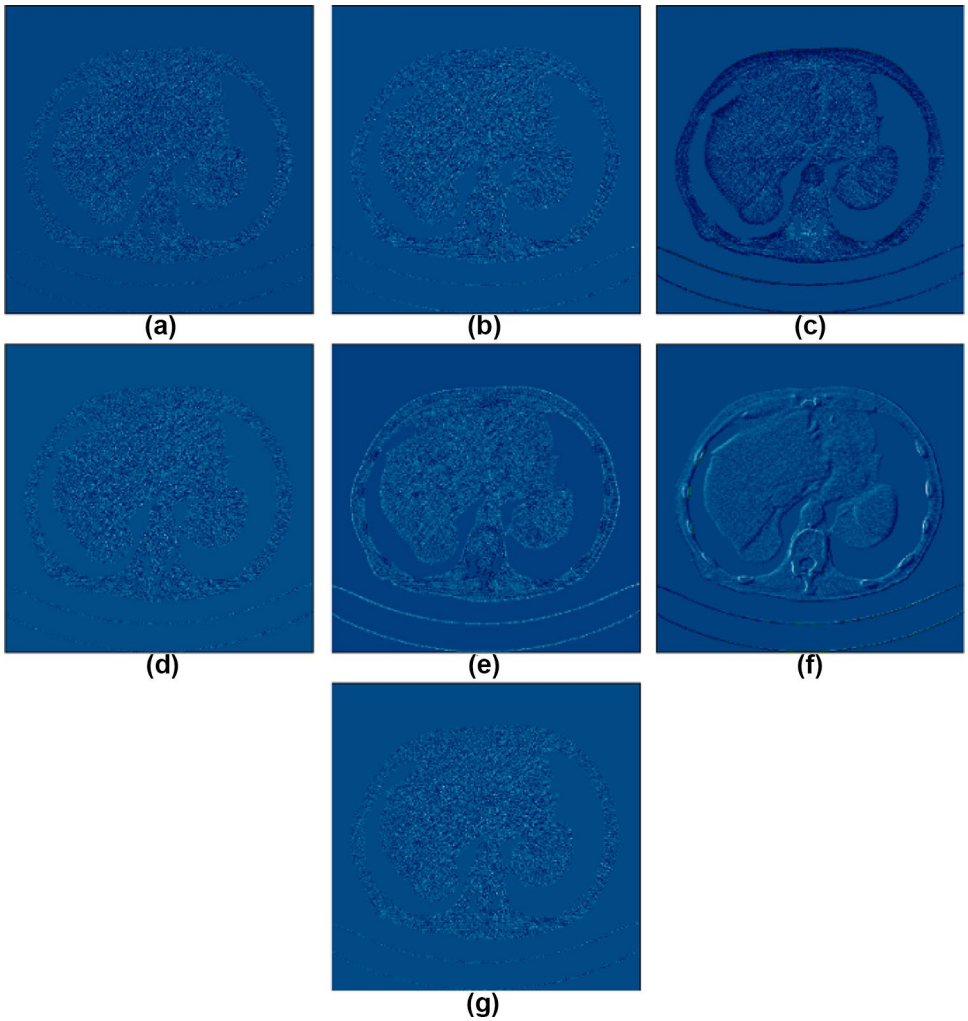

Fig. 10 Same as Fig. 7 but for sample 2

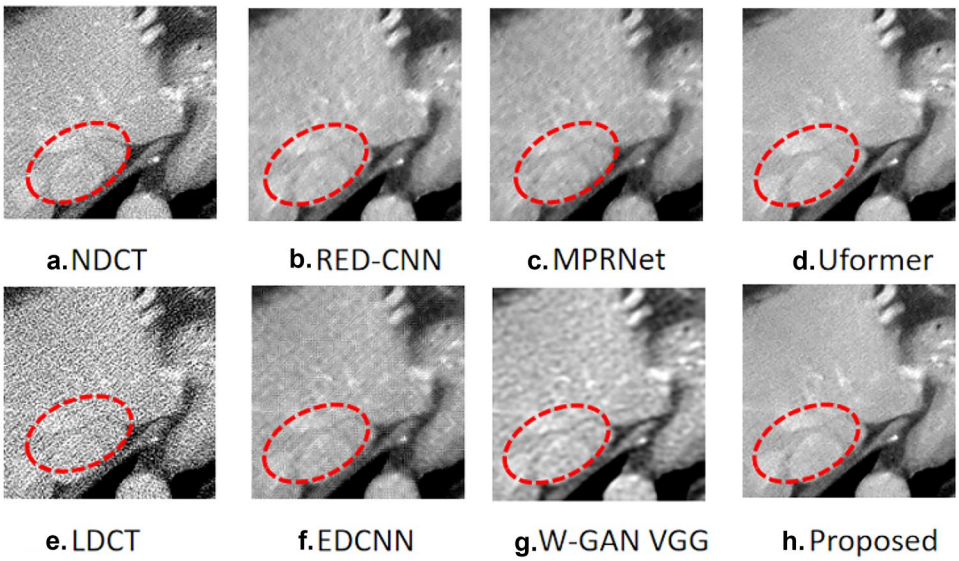

Fig. 11 Same as Fig. 8 but for magnified ROI

details, and the contour of organs is the clearest. While eliminating noise and artifacts, the edge details of the image are better retained, and the quality is close to the corresponding NDCT ground-truth image.

Based on the magnified ROI in Fig. 8, it can be observed that the edge details of the LDCT image with a large amount of quantum noise are excessively fuzzy indicated by the red arrow, and some sharp edge details are lost. Meanwhile, the HCformer has a better visual effect on intuitive noise contrast and is closer to the NDCT image than other methods, as shown in Fig. 6. By comprehensively comparing the magnified ROI of several denoising methods in Fig. 8, the protection of image edges can be realized. By comparing the parts marked by red arrows and blue arrows, the HCformer has clearer edge details and more obvious contrast with the surrounding background.

Figure 9 shows the overall comparison of another representative sample slice in Fig. 5b; Fig. 10 shows the corresponding results of different denoising algorithms, and Fig. 11 shows the comparison of a magnified ROI in Fig. 9. The overall visual comparison in Fig. 9 shows that the HCformer has a fine denoising effect on CT images with different structures. Compared with the magnified ROIs in Fig. 11, we can see that the Transformer-based denoising algorithms also have a good denoising effect on the parts with similar background structures, and they can effectively retain the edge contours of similar parts. The proposed HCformer not only retains more edge details but also has the closest intuitive denoising effect to the NDCT image, and the overall image is cleaner.

The compared results from Figs. 6 to 11 confirm that the HCformer achieves the best visual effect. While removing most of the noise, it can retain the details similar to the traditional normal dose images. Therefore, the proposed algorithm has the best performance in terms of visual examination.

Objective Comparison

To objectively evaluate the performance of the HCformer, the SSIM, HuRMSE, and FSIM are compared with the NDCT images serve as ground-truths.

Table 5 Quantitative comparison of sample 1

	SSIM	HuRMSE	FSIM
LDCT	0.7271	47.4497	0.6858
REDCNN	0.7719	24.7960	0.7161
WGAN	0.7245	28.3132	0.3647
EDCNN	0.7262	27.5753	0.7196
MPR	0.7290	27.7993	0.6912
Uformer	0.7820	24.2618	**0.7181**
HCformer(Ours)	**0.7840**	**24.2136**	0.7172

The bold font indicate that these results are optimal

Table 6 Quantitative comparison of sample 2

	SSIM	HuRMSE	FSIM
LDCT	0.7829	37.7200	0.6973
REDCNN	0.8169	19.9145	0.7247
WGAN	0.7880	26.0794	0.4535
EDCNN	0.7951	25.4540	0.7195
MPR	0.8018	24.5641	0.7015
Uformer	0.8271	19.3647	0.7330
HCformer(Ours)	**0.8304**	**19.3100**	**0.7341**

The bold font indicate that these results are optimal

Table 7 Quantitative comparison of different algorithms in the testing set

	SSIM (Mean ± STD)	HuRMSE (Mean ± STD)	FSIM (Mean ± STD)
LDCT	0.8017 ± 0.0453	34.1898 ± 8.7348	0.6885 ± 0.0284
REDCNN	0.8243 ± 0.0437	20.5948 ± 4.5076	0.6959 ± 0.0249
WGAN	0.8021 ± 0.0567	22.4559 ± 5.2379	0.3692 ± 0.0538
EDCNN	0.8393 ± 0.0388	20.8739 ± 4.5729	0.6974 ± 0.0247
MPR	0.8120 ± 0.0390	19.9183 ± 3.6081	0.6692 ± 0.0246
MEPNet	0.8313 ± 0.03798	19.0477 ± 4.4211	0.7098 ± 0.0257
Uformer	0.8441 ± 0.0409	17.7754 ± 4.3471	**0.7250 ± 0.0253**
HCformer(Ours)	**0.8507 ± 0.0405**	**17.7213 ± 4.3480**	0.7247 ± 0.0255

The bold font indicate that these results are optimal

Tables 5 and 6 summarize the objective evaluation metrics for the representative test samples in Fig. 5, where the best metrics are marked in bold. It can be seen that the proposed HCformer achieves the best performance in both of the cases.

Table 7 shows the comparison of the overall objective evaluation indicators of different algorithms in the testing set. It can be seen that the proposed HCformer has obtained the best results, and the Transformer-based method has achieved better performance in the testing set compared with the CNN convolution network. Compared with the convolutional method, the proposed method has improved the SSIM and similarity of spatial structures. Compared with the UFformer method, the proposed Transformer has achieved the best evaluation index. The HCformer mentioned in this paper also has achieved the best results in the index HuRMSE, which measures the proximity between the testing set and the NDCT images.

Because the results of objective evaluation indexes of Uformer, MEPNet, and MPR are close to those of HCformer, we also performed the independence test for the similar mean values. Specifically, t-test is used to test the mean independence of the evaluation index data of two groups of testing results. The calculation formula is

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{17}$$

where $\overline{X_1}$ is the mean value of the first group of samples, $\overline{X_2}$ is the mean value of the second group of samples, and S_1 and S_2 are the standard deviations of the first and second groups of samples, respectively. The degrees of freedom are

$$v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 v_1} + \frac{s_2^4}{n_2^2 v_2}} \tag{18}$$

where $v_1 = n_1 - 1, v_2 = n_2 - 1$. The testing results are shown in Table 8.

Table 8 Independent probability of objective index of t-test with HCformer, unit(%)

	Uformer(%)	MEPNet(%)	MPR(%)
SSIM	82.64	96.11	99.99
HuRMSE	82.80	98.29	99.99
FSIM	87.41	97.89	99.99

The verification results of the HCformer and UFformer with the closest mean values show that the improved results of HCformer are reliable, and the independence of the two groups of data is also tested. The independence testing results between MEPNet and HCformer show that the probability of independent validation is further improved. Finally, by comparing the experimental results of HCformer and MPR, we can see that they are quite different from each other. Hence, the proposed HCformer is more effective than other algorithms in improving experimental results.

Conclusion

For the LDCT denoising task, a hybrid transformer CNN network (HCformer) is proposed. The network can make use of the advantages of convolution operation and Transformer at the same time to integrate global information and local information. The HCformer network uses the NEF module based on a multidimensional convolution to replace the MLP layer in the standard structure. The combination of depth convolution and per-pixel convolution is used to extract channel-level and pixel-level features, respectively. While the Transformer structure can carry out the wide range correlation of LDCT images, it can also enhance the expression of local information in the network structure to protect the image details. In addition, to avoid the problem that a single MSE loss function will cause the image to be too smooth, this paper uses a combined loss function to constrain the network to pay attention to the edge details in the LDCT image. The proposed HCformer

framework is validated and evaluated on the widely used AAPM data set, and ablation experiments are carried out. The results show that the network can effectively denoise the LDCT images by retaining the detailed information of each tissue structure more clearly. Through objective comparison, it is confirmed that the HCformer outperforms other mainstream algorithms in terms of SSIM, HuRMSE, and FSIM.

Although our proposed algorithm can effectively implement the denoising process for low-dose CT images, there are still potential rooms for further improvement. In the future, we will explore the universal ability of HCformer in multiple sample data, train and test the data on different manufacturers' devices, and improve the generalization ability of network structure. Based on the excellent results in image quality evaluation indicators, we will seek collaboration with clinical radiologists, further improve the practice of the network in clinical applications, and establish a more professional and effective task-driven visual evaluation method.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Luo, L., Hu, Y., & Chen, Y. Research status and prospect for low-dose CT imaging. *J. Data Acquisition Process.* 2015; *30*(1), 224-234.
2. Zhenwei L, & Jianguo Z (2011). Overview of PET/CT image reconstruction technology. *Chinese Journal of Medical Devices*, 35 (1), 53-57
3. Rodrigues, J. C., Pierre, A. F., Hanneman, K., Cabanero, M., Kavanagh, J., Waddell, T. K., ... & Nguyen, E. T. CT-guided microcoil pulmonary nodule localization prior to video-assisted thoracoscopic surgery: diagnostic utility and recurrence-free survival. *Radiology*. 2019; *291*(1), 214–222.
4. Huynh, E., Coroller, T. P., Narayan, V., Agrawal, V., Hou, Y., Romano, J., ... & Aerts, H. J. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiotherapy and Oncology* .2016; *120*(2), 258–266.
5. D.J.Brennerand E.J.Hall, "Computed Tomography as a Source of Increased Radiation Exposure," *New England Journal of Medicine*.2007; Vol. 357, No. 22, 2007, pp. 2277–2284
6. Xiao rong P, Jiaxin H, Junchi L, Jiayu S, Jixiang L, Yue Z,...&Ren Asia (2022). A review of CT image denoising for clinical needs. *Frontier of Data and Computing Development*, 3 (6), 35-49
7. Yungang, Z., Benshun, Y., Chenyue, W., & Yu, F. Low-Dose CT Image Denoising Method Based on Convolutional Neural Network. *Acta Optica Sinica*. 2018;*38*(4), 0410003.
8. Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., & Wang, G. Low-dose CT via convolutional neural network. *Biomedical optics express*. 2017; *8*(2), 679-694.
9. Chen, H., Zhang, Y., Kalra, M. K., Lin, F., Chen, Y., Liao, P., ... & Wang, G. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*. 2017;*36*(12), 2524–2535.
10. Wu, D., Kim, K., Fakhri, G. E., & Li, Q. A cascaded convolutional neural network for X-ray low-dose CT image denoising. 2017; *arXiv preprint* arXiv:1705.04267.
11. Wolterink, J. M., Leiner, T., Viergever, M. A., & Išgum, I. Generative adversarial networks for noise reduction in low-dose CT. *IEEE transactions on medical imaging*. 2017;*36*(12), 2536-2545.
12. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., ... & Wang, G. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*. 2018; *37*(6), 1348–1357.
13. Yi, X., & Babyn, P. Sharpness-aware low-dose CT denoising using conditional generative adversarial network. *Journal of digital imaging*, *31*(5). 2018; 655-669.
14. Yin, X., Zhao, Q., Liu, J., Yang, W., Yang, J., Quan, G., ... & Coatrieux, J. L. Domain progressive 3D residual convolution network to improve low-dose CT imaging. 2019; *IEEE transactions on medical imaging*, *38*(12), 2903–2913.
15. Guo, Z., Zhou, F., Chen, Y., & Yuan, J. A Low-Dose CT Image Denoising Method Combining Multistage Network and Edge Protection. *Tehnički vjesnik*. 2022;*29*(3), 1059-1067.
16. Wang, X., Girshick, R., Gupta, A., & He, K. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018;(pp. 7794–7803).
17. Zhang, H., & Patel, V. M. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018;(pp. 695–704).
18. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. Restormer: Efficient Transformer for High-Resolution Image Restoration. 2021; *arXiv preprint* arXiv:2111.09881.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. 2017;*Advances in neural information processing systems*, *30*.
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020; *arXiv preprint* arXiv:2010.11929.
21. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*. 2019;*32*.
22. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., ... & Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. 2020; *arXiv preprint* arXiv:2006.03677.
23. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*. 2020;(pp. 213–229). Springer, Cham.
24. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. Deep learning for generic object detection: A survey. *International journal of computer vision*. 2020; *128*(2), 261-318.
25. Cordonnier, J. B., Loukas, A., & Jaggi, M. On the relationship between self-attention and convolutional layers. 2019; *arXiv preprint* arXiv:1911.03584.
26. Elsayed, G., Ramachandran, P., Shlens, J., & Kornblith, S. Revisiting spatial invariance with low-rank local connectivity. 2020; In *International Conference on Machine Learning* (pp. 2868–2879). PMLR.
27. Elsayed, G., Ramachandran, P., Shlens, J., & Kornblith, S. Revisiting spatial invariance with low-rank local connectivity. In *International Conference on Machine Learning*. 2020;(pp. 2868–2879). PMLR.
28. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., ... & Gao, W. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021;(pp. 12299–12310).
29. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision . 2021;(pp. 1833–1844).
30. Yang, J., Deng, H., Huang, X., Ni, B., & Xu, Y. Relational learning between multiple pulmonary nodules via deep set attention transformers. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* .2020; (pp. 1875–1878). IEEE.
31. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. 2021; *arXiv preprint arXiv: 2102.04306*.
32. Zhang, Z., Yu, L., Liang, X., Zhao, W., & Xing, L. (2021, September). TransCT: dual-path transformer for low dose computed tomography. *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 55–64). Springer, Cham.
33. QIAO Y Y, QIAO Z W. Low-dose CT image reconstruction method based on CNN and transformer coupling network[J]. *CT Theory and Applications*. 2022, 31(6): 697-707. DOI:https://doi.org/10.15953/j.ctta.2022.114.
34. Luthra, A., Sulakhe, H., Mittal, T., Iyer, A., & Yadav, S. (2021). Eformer: Edge enhancement based transformer for medical image denoising. arXiv preprint arXiv:2109.08044.
35. Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In CVPR, 2021.
36. A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In CVPR, 2005
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. , ... & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021; (pp. 10012–10022).
38. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. Early convolutions help transformers see better. *Advances in0020Neural Information Processing Systems*. 2021;*34*, 30392–30400.
39. Liang T, Jin Y, Li Y, et al. EDCNN: Edge enhancement-based Densely Connected Network with Compound Loss for Low-Dose CT Denoising. arXiv: 2011.00139
40. ZAMIR, SW, ARORA, A., KHAN, S., HAYAT, M. KHAN, F., YANG, M., & SHAO, L. (2021). Multi-stage Progressive Image Restoration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2021, pp. 14821-14831
41. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., & Li, H. . Uformer: A general u-shaped transformer for image restoration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022,pp. 17683–17693).