Nuclear Science and Techniques (2023) 34:61
https://doi.org/10.1007/s41365-023-01208-0

Hformer: highly efficient vision transformer for low-dose CT denoising

Shi-Yu Zhang[1,2,3] · Zhao-Xuan Wang[4] · Hai-Bo Yang[1,2,3] · Yi-Lun Chen[1,2,3] · Yang Li[5] · Quan Pan[4,5] · Hong-Kai Wang[6] · Cheng-Xin Zhao[1,2,3]

Received: 14 December 2022 / Revised: 26 February 2023 / Accepted: 27 February 2023 / Published online: 26 April 2023

Abstract
In this paper, we propose Hformer, a novel supervised learning model for low-dose computer tomography (LDCT) denoising. Hformer combines the strengths of convolutional neural networks for local feature extraction and transformer models for global feature capture. The performance of Hformer was verified and evaluated based on the AAPM-Mayo Clinic LDCT Grand Challenge Dataset. Compared with the former representative state-of-the-art (SOTA) model designs under different architectures, Hformer achieved optimal metrics without requiring a large number of learning parameters, with metrics of 33.4405 PSNR, 8.6956 RMSE, and 0.9163 SSIM. The experiments demonstrated designed Hformer is a SOTA model for noise suppression, structure preservation, and lesion detection.

Keywords Low-dose CT · Deep learning · Medical image · Image denoising · Convolutional neural networks · Self-attention · Residual network · Auto-encoder

1 Introduction

Computed tomography (CT) is a diagnostic imaging method that uses precisely aligned X-rays, gamma-rays, ultrasound, and ion beams [1] to create cross-sectional images of the human body. It uses a highly sensitive detector and focuses X-rays to create 3D images. CT is known for its fast scan time and clear images and is used to examine a variety of diseases. However, it exposes patients to harmful radiation, which may adversely affect their health if the dose is too high.

Low-dose CT (LDCT) has been developed as an alternative to reduce the X-ray dose. LDCT uses less radiation than traditional CT (approximately 1/4 of that of the normal-dose CT) and causes less radioactive damage to the human body. It is particularly suitable for physical examination screening and patients who require multiple examinations. However, unlike NDCT images, LDCT images can also be affected by noise and artifacts in clinical use [2]. Therefore, the suppression of noise and artifacts in LDCT images is an important issue that must be addressed before applying LDCT to clinical diagnosis.

In traditional approaches, researchers use iterative methods to suppress artifacts and noise by relying on physical models and priori information. Unfortunately, these algorithms are difficult to implement in commercial CT scanners

This work was supported by the National Natural Science Foundation of China (Nos. 11975292, 12222512), the CAS "Light of West Chin" Program, the CAS Pioneer Hundred Talent Program, the Guangdong Major Project of Basic and Applied Basic Research (No. 2020B0301030008).

✉ Hai-Bo Yang
yanghaibo@impcas.ac.cn

Hong-Kai Wang
h.wang@cicams.ac.cn

Cheng-Xin Zhao
chengxin.zhao@impcas.ac.cn

[1] Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Advanced Energy Science and Technology Guangdong Laboratory, Huizhou 516003, China

[4] School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China

[5] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[6] Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

because of hardware limitations and high computational costs. With the growing popularity of next-generation artificial intelligence techniques and deep neural networks (DNNs), DNNs have become a mainstream approach to LDCT image denoising, which includes both supervised and unsupervised learning [3]. Recently, most methods have focused on using convolutional neural networks (CNNs) [4, 5] to suppress image noise and have achieved promising results. Although CNNs can learn from large-scale training data and obtain superior solutions, they have limitations in capturing global features in images [6–9] because the pooling layer loses a significant amount of valuable information and ignores the correlation between local and global features. Additionally, typical CNN models lack generic interpretation modules [10]. These deficiencies negatively affect the ability to retrieve richer structural information from denoised images. This also renders the model uninterpretable.

Recently, a transformer model [7] has shown excellent performance in computer vision [11–13] and has been utilized to enhance the quality of LDCT images. Compared with CNNs, transformer models are better at capturing global features and interactions between remote features, thus acquiring richer features in images. In addition, transformer models have a higher visual interpretability owing to their inherent self-attentive block [14, 15]. However, there are two primary limitations to transformer models. First, the complexity of the self-attention mechanism computation is $O(n^2d)$, and excessive computation can cause problems in clinical applications. Second, the transformer is not as adept at extracting local features as CNNs. To address these limitations and better combine the advantages of both CNNs and transformers, this study proposes the Hformer module, which combines the advantages of vision transformers to achieve a lighter structure and improved results. Specifically, Hformer comprises the following two aspects:

A more lightweight convolution encoder. The convolution module consists of multiple 3×3 depthwise separable convolution (DSC) blocks. Depthwise convolution has a relatively low number of computational parameters. It applies one convolution kernel to each channel of the input feature map and then combines the outputs of all convolution kernels to obtain its final output. The number of output channels for the convolution operation is equal to the number of convolution kernels, and only one convolution kernel is used per channel in depthwise convolution. Therefore, the number of output channels for a single channel after the convolution operation is also one. In this study, two depth-separable convolution layers were used to enrich the local representation, whereas standard layer normalization (LN) and the Gaussian error linear unit (GELU) were used to activate nonlinear feature mapping. Finally, a skip connection was added to allow information to flow through the network hierarchy. This block is similar to the ConvNeXt block but with a smaller kernel size to promote a more lightweight model.

More efficient patch-based global interactions encoding module. The self-attention module was suitable for learning global representations. Understanding the intrinsic features of visual tasks is crucial. To take advantage of this, while minimizing the model overhead, we use cross-covariance attention to integrate attention operations on the channel features instead of using attention operations on the global features in the feature map. This approach effectively reduces the complexity of the self-attention operation, thereby reducing the computational time from $(HW)^2C$ to HWC^2, which is about the linear relationship of image resolution. This method not only reduces the computational effort from quadratic with respect to the image resolution but also effectively and implicitly encodes local contextual information.

2 Related works

2.1 Traditional

LDCT image denoising is a research area with important clinical applications in medical image denoising. In the early years, researchers mainly used preprocessing methods such as iterative reconstruction (IR)-based algorithms for denoising LDCT images. This method combines the statistical properties of the data in the sinogram domain, prior information in the image domain, and parameters of the imaging system into a unified objective function. Using compressive sensing (CS) [16], some image priors are represented as sparse transforms to deal with low-dose, few-view, finite-angle, and internal CT problems, such as full variational (TV) and its variants [17], non-local averaging (NLM) [18], dictionary learning [19], and low-ranking [20]. Although IR methods have achieved promising results, they have two limitations. First, the IR techniques are less scalable and migratory. Because this technique needs to be preconfigured for a specified device, users and other vendors do not have access to detailed information about the specific scanner geometries and calibration steps. Second, the computational overhead associated with the information retrieval techniques is significant. This poses a significant challenge in clinical applications.

Another option is to post-process the reconstructed LDCT image, which does not depend on the original image and can be applied directly to LDCT images without the need for preset modules in any CT system. Li et al. [21] used the NLM to reconstruct feature similarities within large neighborhoods in images. Inspired by sparse representation theory, Aharon et al. applied dictionary learning [22] to denoise LDCT images and significantly improved the denoising quality in

the reconstruction of abdominal images [23]. Feruglio et al. demonstrated that block-matching 3D (BM3D) is effective for various X-ray imaging tasks [24]. However, unlike the other two methods, this method does not accurately determine the noise distribution in the image domain, which hinders the user from achieving the best compromise between structure preservation and noise substitution. In general, the accuracy of these traditional methods remains low, owing to data volume limitations [25].

2.2 Deep learning based methods

Efficient data-driven deep learning methods have great potential in intelligent medicine owing to the limitations of data volume and the consequent low accuracy of traditional methods. It has achieved promising results in various applications such as lesion classification, image quality improvement, and organ segmentation. Deep learning can mimic human information processing by efficiently learning high-level features from pixel data through a hierarchical network framework. Thus, it has been widely used for LDCT image reconstruction. In general, deep-learning-based LDCT image denoising methods can be divided into three categories: convolutional neural network (CNNs)-based methods, transformer-based methods, and their combination.

2.2.1 CNN in LDCT

Researchers have used CNN network-based methods to denoise LDCT images. For example, Chen et al. [26] applied lightweight CNNs to an LDCT imaging framework and obtained preliminary results. Wurlf et al. [27] mapped the filtered back projection (FBP) workflow to a deep CNN architecture to reduce the reconstruction error to 1/2 of its original value in the case of limited-angle laminar imaging. Chen et al. [28] proposed the REDCNN model, which utilizes convolution, deconvolution, and shortcut connections to construct residual coding and decoding convolutional neural networks that have been well evaluated for noise suppression, structure preservation, and lesion detection. Chen et al. [29] proposed the NCS-Unet model, in which the exceptional characteristics of the non-subsampled contourlet transform (NSCT) and Sobel filter are introduced into NCS-Unet. NSCT effectively separates convolved features into high- and low-frequency components, which allows the strengths of both types of information to be merged. Liu et al. [30] proposed a 3D residual convolutional network to iteratively estimate the reconstructed images from the LDCT resolution. Their method avoids time-consuming iterative reconstructions. Ma et al. [31] implemented an attention-residual dense convolutional neural network (CNN) approach, referred to as AttRDN. The AttRDN approach employs an attention mechanism that combines feature fusion and global residual learning to remove noise from contaminated LDCT sinograms effectively. The denoising process was achieved by first extracting noise from the noisy sinogram using the attention mechanism and then subtracting the noise obtained from the input sinogram to restore the denoised sinogram. Finally, the CT image was reconstructed using filtered back projection. Xia et al. [32] proposed a framework called the parameter-dependent framework (PDF), which facilitates the simultaneous training of data with various scanning geometries and dose levels. In the proposed framework, the scanning geometry and dose level are parameterized and input into two multilayer perceptrons (MLPs). These MLPs are utilized to regulate the feature maps of a CT reconstruction network, thereby conditioning the network outputs on different scanning geometries and dose levels. Lu et al. [33] presented a pioneering investigation into the application of a neural architecture search (NAS) to LDCT, which culminated in the development of a memory-efficient multiscale and multilevel NAS solution named M3NAS. M3NAS synthesizes features from various scale cells to detect multiscale structural details in the image while searching for a hybrid cell and network-level structure to optimize the performance. M3NAS also substantially reduces model parameters and enhances inference speeds. Huang et al. [34] proposed a two-stage residual CNN, where the first stage uses a smooth wavelet transform for texture denoising and the second stage combines the mean wavelet transform to enhance image structure. Tan et al. [35] proposed a new method for reducing noise in LDCT images using a selective feature network and the unsupervised learning model, CycleGAN. This approach adaptively selects features to enhance image quality. Despite the interesting results of CNNs for LDCT, CNN-based models typically lack the ability to capture global contextual information owing to the characteristics of the limited sensory field of CNNs and, thus, are less efficient in modeling the structural similarity of the entire image [36].

2.2.2 Transformer in LDCT

In recent years, the transformer-based architectures pioneered by Dosovitskiy et al. [37], which successfully exploited transformers for image classification tasks, have achieved great success in the field of computer vision. Since then, several transformer-based models have been used to solve downstream vision tasks with excellent results, including image super-resolution [11], denoising [38], and colorization [39]. In LDCT image denoising, Wang et al. [40] designed a Uformer with the ability to capture useful dependencies for image restoration using non-overlapping window-based self-attentiveness to reduce computational effort while employing deep convolution in the forward network to further improve its ability to capture the local

context. They achieved excellent results in multiple image restoration tasks (e.g., image noise reduction, image rain removal, and image deblurring). Luthra et al. [41] combined the learnable Sobel-Feldman operator for edge enhancement and built a transformer architecture-based codec network, Eformer, for medical image denoising, based on the self-attentive mechanism of non-overlapping windows. Wang et al. [42] used a more powerful token-rearranged replacement convolutional neural networks to include local contextual information and proposed a convolution-free Token2Token dilated vision transformer (CTformer) for LDCT image denoising.

2.2.3 Combination of transformer and CNN

Self-attention is widely used in CNNs for visual tasks. The primary research direction is to combine VIT and CNNs to design new backbones. Graham et al. [43] mixed convnet and transformer in their LeVit model, and LeVit significantly outperformed the previous convnet and ViT models in terms of the speed and accuracy tradeoff. Zhang et al. [44] combined the local modeling capability of the residual convolution layer with the non-local modeling capability of the Swin transformer block and then inserted them into the UNet architecture as the main building block to achieve outstanding results in image noise reduction. CoatNet [45] combines convolution and self-attention to design a novel transformer module that allows the model to focus on more local and global information simultaneously. Another idea is to modify the transformer block using convolution, such as replacing the multiheaded attention with a convolutional layer [46], adding additional convolutional layers in parallel [47] or serially [48] to capture local relations. In addition, some researchers have used local transformer modules in convolution-based network architectures to enhance access to global information. For example, Srinivas [49] proposed a simple but powerful backbone architecture, BoTNet, which simply replaces spatial convolution with global self-attention in the last three bottleneck blocks of ResNet and achieves strong performance in image recognition. ConViT [50] integrates soft convolutional induction bias through gated positional self-attention. The CMT [51] block comprises a deep convolution-based local perceptual unit and a lightweight transformer module.

We found that these hybrid network structures combining convnet and transformer are similar in terms of design ideas. They use convnet to extract local feature information and self-attention to extract local contextual information. Inspired by these works, we integrated the advantages of both CNN and transformer architectures efficiently, and our work helped us achieve SOTA results on LDCT image denoising.

3 Methods

3.1 Denoising model

Our study started from CT images obtained from low-dose scan data reconstructed by filtered back projection (FBP). The noise distribution in CT images typically includes a combination of quantum Poisson and electron Gaussian noises. However, the reconstructed images always have a complex and uneven noise distribution. Furthermore, there is no accurate mathematical model that can describe the relationship between NDCT and LDCT. This makes obtaining high-quality denoising results for LDCT images using traditional methods challenging.

Therefore, the noise distribution can be more accurately modeled using deep learning methods because deep learning is independent of the statistical distribution of image noise. LDCT image denoising can be simplified to address the following problems. Assuming $x \in R^{N\times N}$ represents the LDCT image and $y \in R^{N\times N}$ represents the corresponding NDCT, our goal is to identify a function F that maps from x to y:

$$y = F(x), \tag{1}$$

where $F : R^{N\times N} \rightarrow R^{N\times N}$ indicates a process involving the elimination of image noise and artifacts.

3.2 Network architecture

As shown in Fig. 1, our network uses a self-encoder structure for residual learning that includes two convolutional layers, three Hformer blocks, and four scale layers. The scale layer has a residual connection between 2×2-strided convolution-based down-sampling and 2×2-transposed convolution-based up-sampling. Within the encoder, the down-sampling module employs convolution to reduce the patch size while simultaneously increasing the number of channels between each level. In contrast, the up-sampling module within the decoder utilizes transposed convolution to increase the patch size while concurrently reducing the number of channels between each level. This structure is not only suitable for supervised learning of noise distribution but also for image reconstruction and denoising tasks. Next, we present the details of our study.

3.2.1 Autoencoder

An autoencoder (AE) was originally developed for supervised feature learning based on noisy inputs and is also applicable to image reconstruction. Both CNNs and transformers have shown excellent performance in image denoising. However, because CNNs use local perceptual fields for

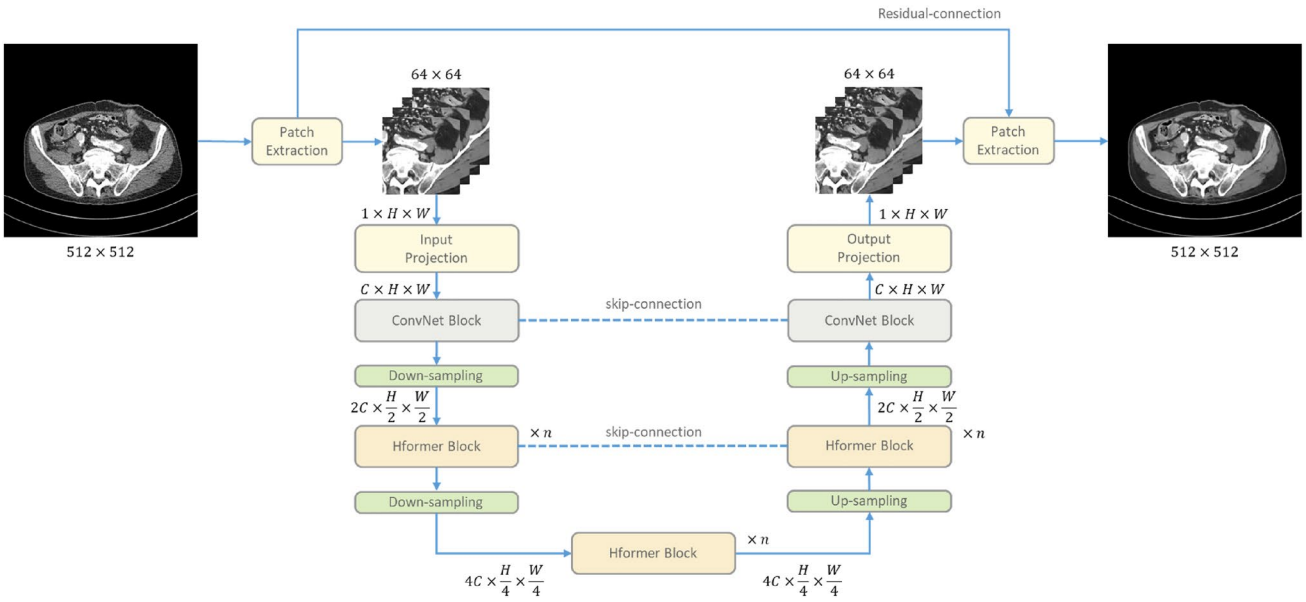

Fig. 1 Overall architecture of Hformer

feature capture, they cannot directly model global environments. The transformer compensates for this deficiency. Therefore, for LDCT, we propose a residual network combining three novel technologies, namely AE, CNNs, and transformers, which originated from the work [52]. Instead of using fully connected layers for encoding and decoding, we performed feature extraction and image reconstruction symmetrically. Moreover, unlike typical encoding structures, it includes residual learning with shortcuts [4] to facilitate the operation of a shallow information-focused convolutional layer and the corresponding deconvolutional layer. In addition, this approach solves the gradient disappearance problem, such that deep models can be stably trained [53].

3.2.2 Patch extraction

The training process of deep-learning models requires a large number of samples. However, this requirement is often not easily satisfied in practice with adequate samples, especially for medical imaging. In this study, we used overlapping slices in the CT images. This strategy has been shown to be effective in previous studies, where more slices allow the model to detect perceived differences in local areas and significantly increase the number of samples [54]. In our experiments, we extracted fixed-size patches from LDCT images and the corresponding NDCT images.

3.2.3 Residual learning

The convolution operation gradually extracts information from the underlying features to the highly abstract features. The deeper the network, the more abstract (semantic) features that can be extracted. For traditional convolutional neural networks, simply increasing the depth of the network can easily result in gradient disappearance and explosion. Common solutions to this issue include normalized initialization and intermediate normalization layers. However, this leads to the problem of network degradation, which means that as the number of layers in the network increases, the accuracy of the training dataset saturates or even decreases as the number of layers increases. This phenomenon is different from and overfitting does not show a decrease in the accuracy of the training set.

It is common sense that the solution space of the deeper network structure contains the solution space of the shallow network structure, which means that the deeper network structure can obtain better solutions and perform better than the shallow network. However, this is not the case because deeper networks may have worse training and testing errors than shallow networks. This proves that it is not due to overfitting. This phenomenon is probably caused by the stochastic gradient descent strategy and the complex structure of the deep network, which does not result in a globally optimal solution but rather a locally optimal solution.

Therefore, residual learning provides a new way of thinking: since deep networks have degeneracy problems compared to shallow networks, is it possible to retain the depth of deep networks and have the advantage of shallow networks to avoid degeneracy problems? If the later layers of the deep network are learned as a constant mapping $h(x) = x$, the model degenerates into a shallow network. However, it is often difficult to directly learn

this constant mapping. Therefore, we require a different approach: we redesign the network into a new form: $H(x) = F(x) + x \rightarrow F(x) = H(x) - x$. As long as $F(x) = 0$, this constitutes a constant mapping $H(x) = x$, where $F(x)$ is the residual.

Residual learning provides two methods for solving the degradation problem: identity and residual mapping. The residual learning structure is implemented using a forward neural network and shortcut linkage, where the shortcut linkage is equivalent to simply performing the same mapping without generating additional parameters or increasing the computational complexity. The entire network can be trained using end-to-end backpropagation.

Therefore, residual learning is used to avoid the problem of gradient disappearance. This allows the deep model to be trained stably.

3.2.4 Convolution block

Considering that shallow information contains more detailed information (contour, edge, color, texture, and shape features), using CNNs to extract features by sharing convolutional kernels ensures a reduced number of network parameters and improves model efficiency. CNNs exhibit two inherent inductive biases: translational invariance and local correlation. This feature allows CNNs to capture additional local information. Inspired by this, we designed a shallow feature extraction (reconstruction) module consisting primarily of depth-separable convolutions [55]. The feature layer is normalized after a depth-separable convolution and combined with the normalization of the standard layer [56]. Then, two projection convolutions are used to enhance the local representation and channel dimension transformation: A Gaussian error linear unit [57] (GELU) is connected after the first projection convolution to activate it for nonlinear feature mapping. Finally, a residual join is used to smooth the back-and-forth propagation of the information. This process can be formulated as Eq. (2), and its architecture is shown in Fig. 2.

$$x_{i+1} = x_i + \text{Linear}(\text{GeLU}(\text{Linear}(\text{LN}(\text{DWConv}_{7\times7}(x_i))))) \tag{2}$$

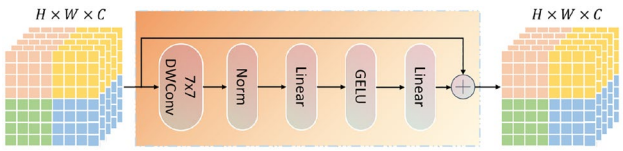

Fig. 2 (Color online) Architecture of convolution block

3.2.5 Hformer block

The Hformer block proposed in this study consists of a depth-wise convolution (DWConv)-based perceptual module and a transformer module with a lightweight self-attentive (LSA) module, as shown in Fig. 3. These two modules are described in detail below.

DWConv based perceptual module. To compensate for the loss in the image domain, we used DWConv with a kernel size of 7×7 in the convolutional perception module to

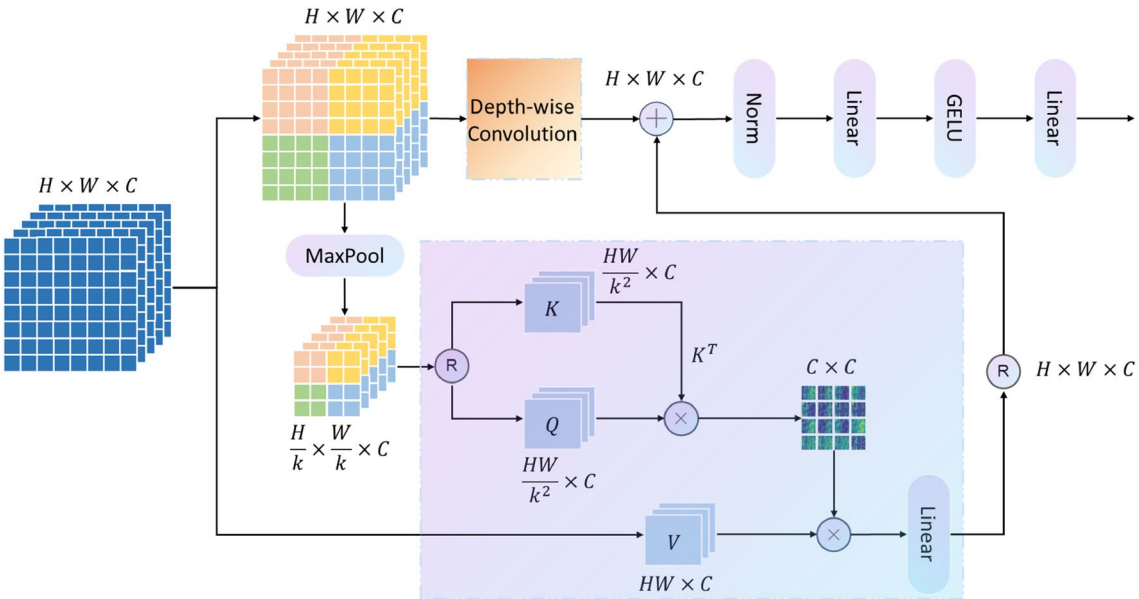

Fig. 3 (Color online) The structure of Hformer block

process the input features and extract features from the local perceptual field in the same manner as a conventional convolution. This approach was inspired by the fact that there are many similarities between local self-attention and DWConv. First, the latter also has sparse connectivity; that is, the computation exists only within the kernel size, and there is no connection between individual channels. DWConv also exhibited weight-sharing properties. However, convolution kernels are shared across all spatial locations and different channels use different convolution kernels, which significantly reduces the number of parameters. In addition, the DWConv kernel is a scientific training parameter that is fixed once training is completed, whereas the computation of attention is a dynamic process. Local self-attention requires positional coding to compensate for the lost positional information, whereas DWConv does not.

Light-weight self-attention. The transformer's original self-attention has a huge overhead, which is a huge burden on computational power. To address this difficulty and obtain valid local contextual information, we reduced the dimensionality of the feature map in our Hformer module and attempted to compute the attention in the channel dimension. Given an input $X \in R^{H\times W\times C}$, the original self-attentive mechanism first generates the corresponding query (Q), key (K), and value (V) (of the same size as the original input) and then generates a weight matrix of size $R^{N\times N}$ through the dot product of Q and K.

$$\begin{cases} Q = W^Q X \\ K = W^K X \\ V = W^V X \end{cases} \tag{3}$$

where W^Q, W^K and W^V are the linear operations. Previous self-attention calculations were performed along the spatial dimension between Q and K, and the results are as follows:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V, \tag{4}$$

where the scaling factor $\frac{1}{\sqrt{d_k}}$ is based on network depth. However, this process usually requires large computational capacity (video memory) owing to the large size of the input features, which makes it difficult to train and deploy the network. Therefore, we used the maximum pooling method to downsample the generation of K and Q separately to obtain two relatively small features, K' and Q':

$$K' = \text{Maxpool}(K) \in R^{\frac{HW}{k^2}\times C}, \tag{5}$$

$$Q' = \text{Maxpool}(Q) \in R^{\frac{HW}{k^2}\times C}. \tag{6}$$

To further reduce the overhead of the model and the algorithm complexity to a linear relationship with the image resolution, we used the attention computed on the channel dimension to implicitly encode patch-based global interactions.

We transpose K' and apply the dot product to K'^{T} and Q' in the channel dimension, and the computed results are supplemented with Softmax to obtain the attention score matrix $\text{Attn}_{\text{channels}}$ with dimension $C \times C$, which is applied to V and obtain the final attention map. The computational effort of this step is $C^2(HW)$, which is linear in image resolution and substantially reduces complexity. The attention operation for channel dimensions can be expressed as follows:

$$\text{Attn}_{\text{channels}}(Q', K', V') = \text{softmax}\left(\frac{K'^T Q'}{\sqrt{d_k}}\right)V \tag{7}$$

4 Experiment

Dataset. We used the publicly released clinical dataset from the 2016 NIH-AAPM Mayo Clinic LDCT Grand Challenge [58] for model training and testing. The dataset consisted of 2378 low-dose (quarter) images and 2378 normal-dose (full) CT images from 10 anonymous patients with 3.0-mm whole-layer slices. We selected patient L506 data for testing, which contained 211 slice images numbered from 000 to 210. We used the data from the remaining nine patients for model training.

Model training and optimization. Our network is an end-to-end mapping M from LDCT images to NDCT images. For the given training data $P = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ where X_i and Y_i denote LDCT and NDCT image patches, respectively, n is the total number of training samples. The model performance can be improved by minimizing the loss $L(X, \theta)$ between the output CT image and the reference NDCT image, where θ refers to learnable parameters. This process can be achieved by optimizing the mean square error (MSE) loss function, as shown in Eq. (8).

$$L(\theta) = \frac{1}{N}||Y_i - M(X_i)|| \tag{8}$$

Experiment setup. The experiments were run on CentOS 7.5 with an Intel Xeon Scalable Gold 6240 CPU@2.6 GHz, using PyTorch 1.11.0, and CUDA 11.2.0. The model was trained using eight NVIDIA Tesla V100 32GB GPU HBM2. For each image, four blocks randomly extracted from all available slices were used for training. The batch size is 16 through 4000 epochs. The ADAM-W optimizer was used to minimize the mean squared error loss, and the learning rate was 1.0×10^{-5}.

4.1 Denoising performance

The performance of our net was compared with other SOTA models, such as RED-CNN [28], SCUNet [44] Uformer [40], DU-GAN [59], and CTformer [42]. The selected models were popular LDCT or natural image denoising models published in top journals and conferences. SCUNet and Uformer are mainstream deep learning-based image-noise reduction algorithms. Red-CNN is the masterpiece of the convolutional neural network-based CT noise reduction algorithm, and CTformer is the most advanced noise reduction algorithm based on the LDCT dataset, which has excellent results in image noise reduction tasks. We retrained all the models based on their officially disclosed codes.

For quantitative evaluation, we selected the root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) as the quantitative evaluation metrics for image quality. RMSE is a measure of accuracy that can be used to compare the predictive performance of different models on the same dataset and can magnify the error magnitude between the reconstructed image and the ground truth image (the larger the error the larger the RMSE). This representation is shown in Eq. (9):

$$\mathrm{RMSE} = \frac{1}{m}\sum_{i=1}^{m}(\mathrm{im} - \mathrm{gt})^2. \tag{9}$$

PSNR provides an objective criterion for describing the level of image distortion and noise (shown in Eq. 10). The larger the value, the smaller the difference between the reconstructed and reference images.

$$\begin{aligned}\mathrm{PSNR} &= 10 \times \log_{10}\left(\frac{(2^n-1)^2}{\mathrm{MSE}}\right)\\ \mathrm{MSE} &= \frac{1}{m}\sum_{i=1}^{m}(\mathrm{im} - \mathrm{gt})^2\end{aligned} \tag{10}$$

SSIM evaluates the similarity of two images in three ways, and SSIM is defined as Eq. (11).

$$\begin{aligned}\mathrm{SSIM}(\mathrm{im},\mathrm{gt}) &= L(\mathrm{im},\mathrm{gt})C(\mathrm{im},\mathrm{gt})S(\mathrm{im},\mathrm{gt})\\ L(\mathrm{im},\mathrm{gt}) &= \frac{2\mu_{\mathrm{im}}\mu_{\mathrm{gt}} + c_1}{\mu_{\mathrm{im}}^2 + \mu_{\mathrm{gt}}^2 + c_1}\\ C(\mathrm{im},\mathrm{gt}) &= \frac{2\Sigma_{\mathrm{im}}\Sigma_{\mathrm{gt}} + c_2}{\Sigma_{\mathrm{im}}^2 + \Sigma_{\mathrm{gt}}^2 + c_2}\\ S(\mathrm{im},\mathrm{gt}) &= \frac{\Sigma_{\mathrm{im,\,gt}} + c_3}{\Sigma_{\mathrm{im}}\Sigma_{\mathrm{gt}} + c_3}\end{aligned} \tag{11}$$

where μ_{im} and μ_{im}^2 are the mean and variance of the reconstructed image, respectively; μ_{gt} and μ_{gt}^2 are the mean and variance of the ground truth image, respectively; $\Sigma_{\mathrm{im,\,gt}}$ is the covariance between the reconstructed and ground truth images; c_1, c_2, and c_3 are constants. The structural similarity index measures the degree of image distortion and the degree of similarity between two images. Unlike MSE and PSNR, which measure the absolute error, SSIM is a perceptual model, that is, it is more in line with the intuition of human eyes. Its value ranges from zero to one. The higher the value of SSIM, the higher the similarity between the reconstructed and ground truth images. The number of trainable parameters (Param) was used to evaluate model complexity. Table 1 lists the average metrics of all models for L506 patients. Our model has the lowest average RMSE among the SOTA methods. This indicates that our model effectively suppresses noise and artifacts and maintains a high degree of spatial smoothing. In terms of information reconstruction, our model has the best SSIM compared to its competitors, preserving the structural details of the reconstructed images. Meanwhile, CTformer had fewer trainable parameters than ours. Therefore, we conclude that our network is the best noise eliminator compared to its competitors.

4.2 Visual evaluation

To evaluate the denoising ability of the Hformer proposed in this study with the above comparison method, we provided slices 034 and 057, two representative results from a test set consisting of L506 patient data and their corresponding ROI images. The results are shown in Figs. 4, 5, 6 and 7. The corresponding metrics are listed in Tables 2 and 3. Figures 4 and 6 show the results of the abdominal CT images. The noise shown in Fig. 4a is primarily distributed within the abdomen. The outline of the organs and details of the tissue structure were significantly by noise. Obvious streaking artifacts can be observed in the spine and liver, which greatly affect the clinical diagnosis of lesion areas. It is easy to see that convolutional network-based RED-CNN effectively eliminates most of the noise and artifacts and is better at retaining the details.

However, RED-CNN is less effective in the structural recovery of images because it has computational characteristics that can extract high-frequency information more effectively, such as image texture details. Moreover, RED-CNN is limited by the size of the perceptual field and cannot effectively extract more global information. From the results, we can observe that there is over-smoothing of the detailed textures in Uformer and CTformer. This is due to the lack of a convolution layer, which results in blurred CT images.

For noise reduction and the ability to retain detailed structures, the Hformer proposed in this paper also outperforms SCUNet. The denoising performance in the liver and lesion regions in Fig. 4f is significantly better than that in Fig. 4c. Compared with SCUNet based on a parallel structure combined with convolution and self-attention, Hformer based on

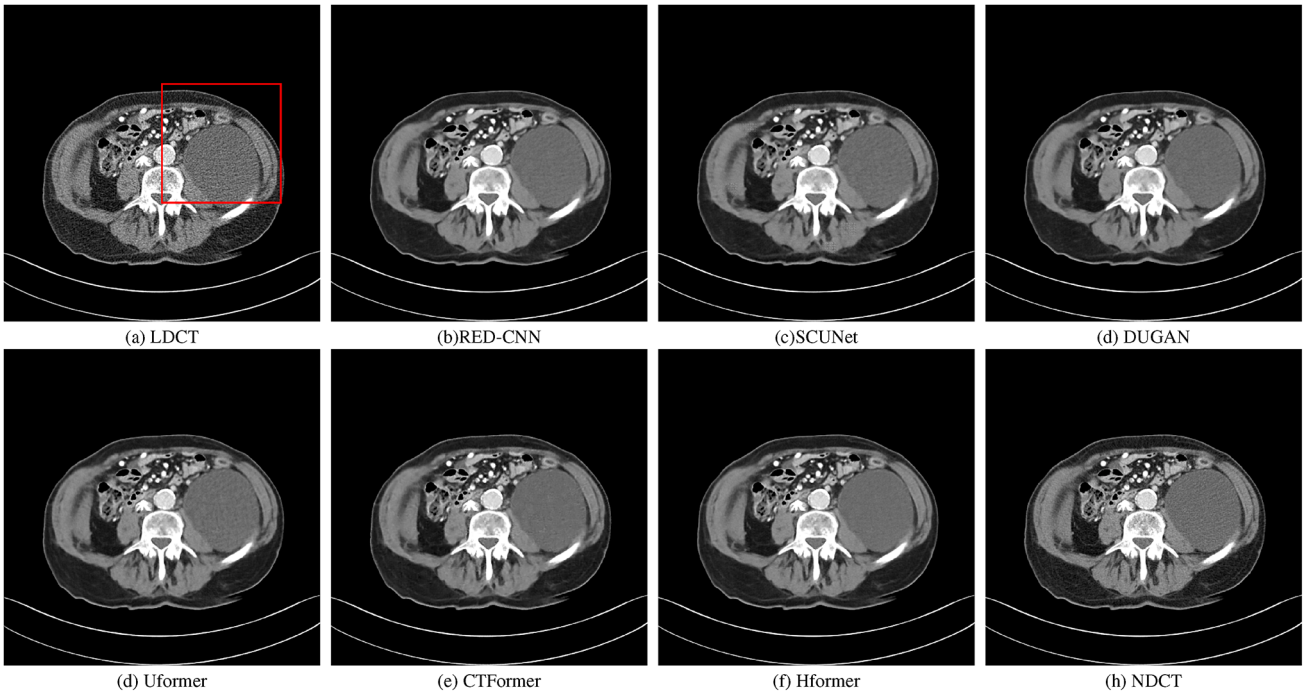

Fig. 4 Results of abdominal slice L506-034 from the testing set using different methods. The display window ranges from – 160 to 240 HU

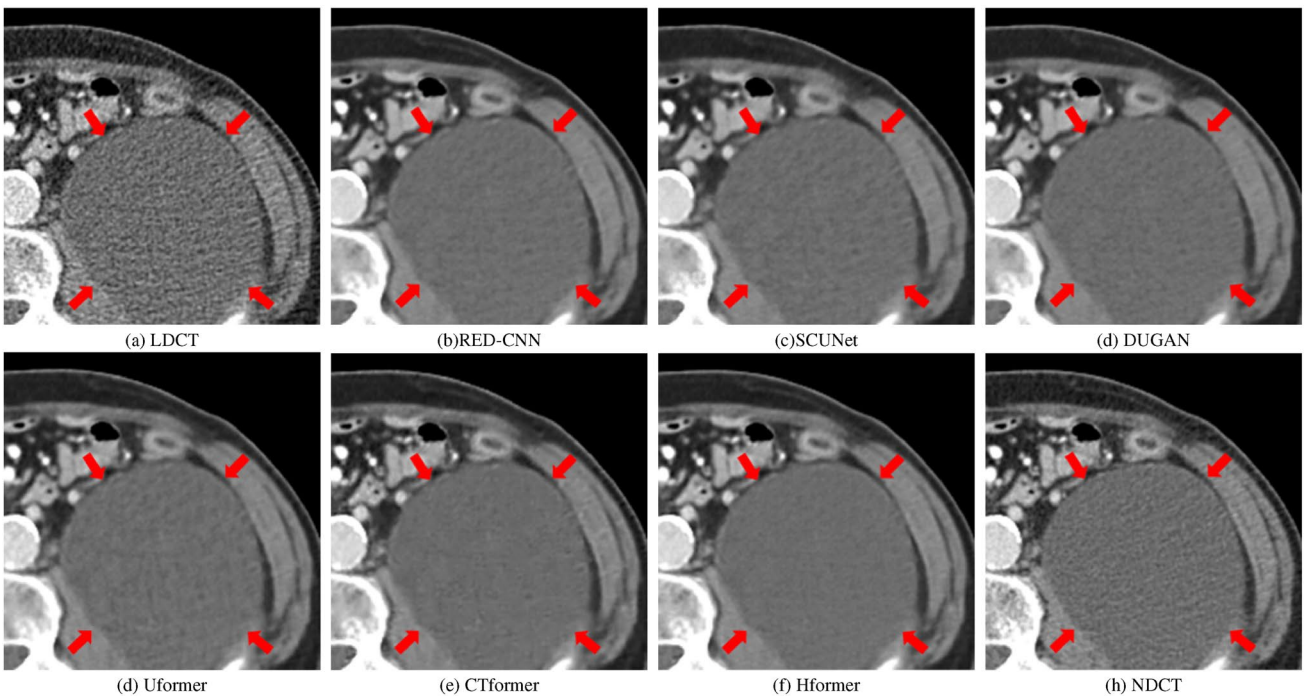

Fig. 5 The corresponding ROI of Fig. 4

a multiscale convolution module and lightweight self-attention exhibits stronger generalization ability and is superior in reconstructing LDCT.

To further demonstrate the performance of Hformer, we provide a magnified image of the ROI marked with a rectangular dashed line in Fig. 4, as shown in Fig. 5. The

Table 1 Quantitative evaluation results of different methods on L506 using the number of learnable parameters (#param.), RMSE, SSIM, and PSNR. Our results are the bold-faced numbers

	#param. (M)	RMSE	SSIM	PSNR
LDCT	–	14.2416	0.8759	29.2489
SCUNet	13	9.4381	0.9066	32.6993
Uformer	12	9.3102	0.9106	33.0623
RED-CNN	1.85	9.0664	0.9109	33.0695
CTformer	1.45	9.0233	0.9121	33.0952
DU-GAN	114.61	8.9464	0.9118	33.1859
Hformer	**1.65**	**8.6956**	**0.9163**	**33.4405**

Table 2 Quantitative results of patient L506's abdominal section 034

Network	RMSE	SSIM	PSNR
LDCT	12.1360	0.8804	30.3597
SCUNet	8.4252	0.9126	33.5296
Uformer	8.0657	0.9193	33.9083
RED-CNN	8.0850	0.9172	33.8876
CTformer	7.9236	0.9190	34.0627
DU-GAN	7.9519	0.9181	34.0318
Hformer	**7.6457**	**0.9235**	**34.3729**

Table 3 Quantitative results of patient L506's abdominal section 057

Network	RMSE	SSIM	PSNR
LDCT	16.2190	0.8424	27.8407
SCUNet	10.3276	0.8821	31.7612
Uformer	10.3909	0.8842	31.7081
RED-CNN	10.0407	0.8859	32.0059
CTformer	10.1807	0.8835	31.8857
DU-GAN	9.9153	0.8866	32.1151
Hformer	**9.7170**	**0.8915**	**32.2906**

arrow-marked region is a piece of tissue with a uniform density distribution. However, almost none of the other methods, except Hformer and CTformer correctly reconstructed the internal details of the lesion region. SCUNet, Uformer, RED-CNN, and CTformer introduced more noise into the image, making it difficult to distinguish the density distribution of this tissue. In our study, DU-GAN and the proposed Hformer were effective in recovering the details and overall structure, and Hformer performed better than DU-GAN in suppressing artifacts.

Another result for the test set is shown in Fig. 6, and its ROI is shown in Fig. 7. Owing to the reduced radiation dose, the structures of many soft tissues are more affected by noise during reconstruction. The internal details of organs are difficult to distinguish accurately. Although Uformer and SCUNet reconstructed the organ contours well, and the organ boundaries were clearly visible, a large amount of noise was generated inside the organ. As shown in Fig. 7, only Hformer

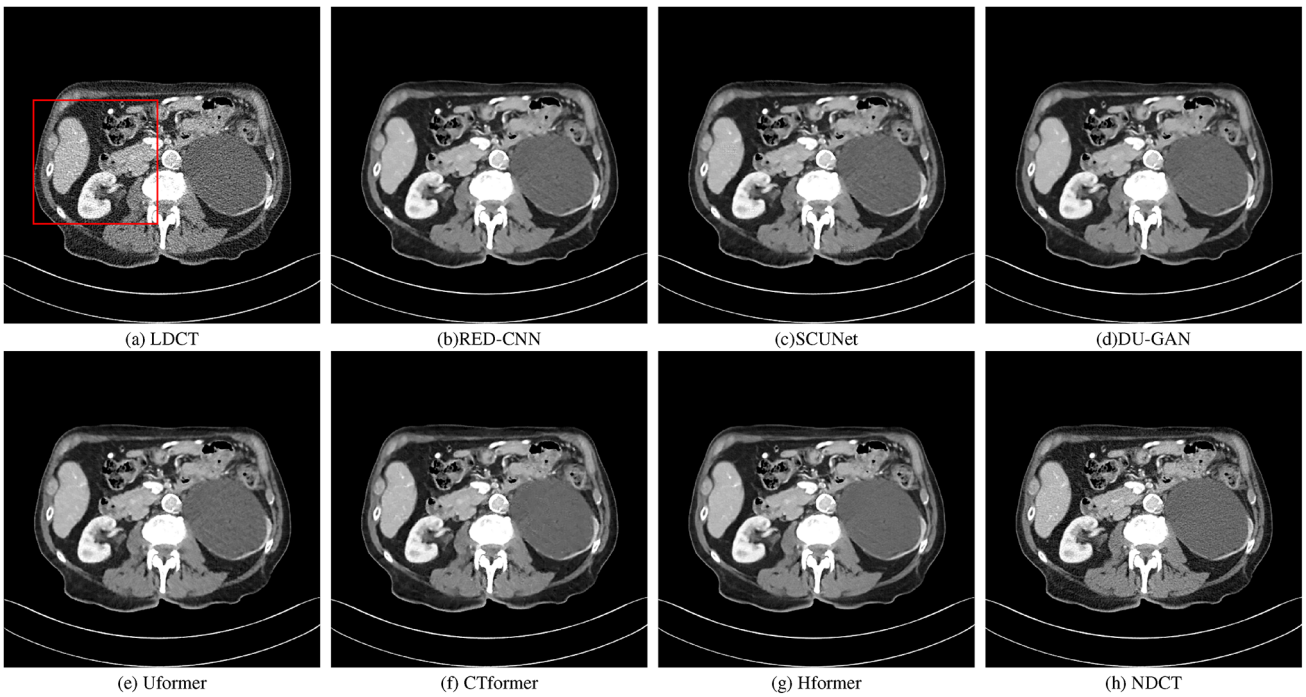

Fig. 6 Results of abdominal slice L506-057 from the testing set using different methods. a low dose. The display window ranges from – 160 to 240 HU

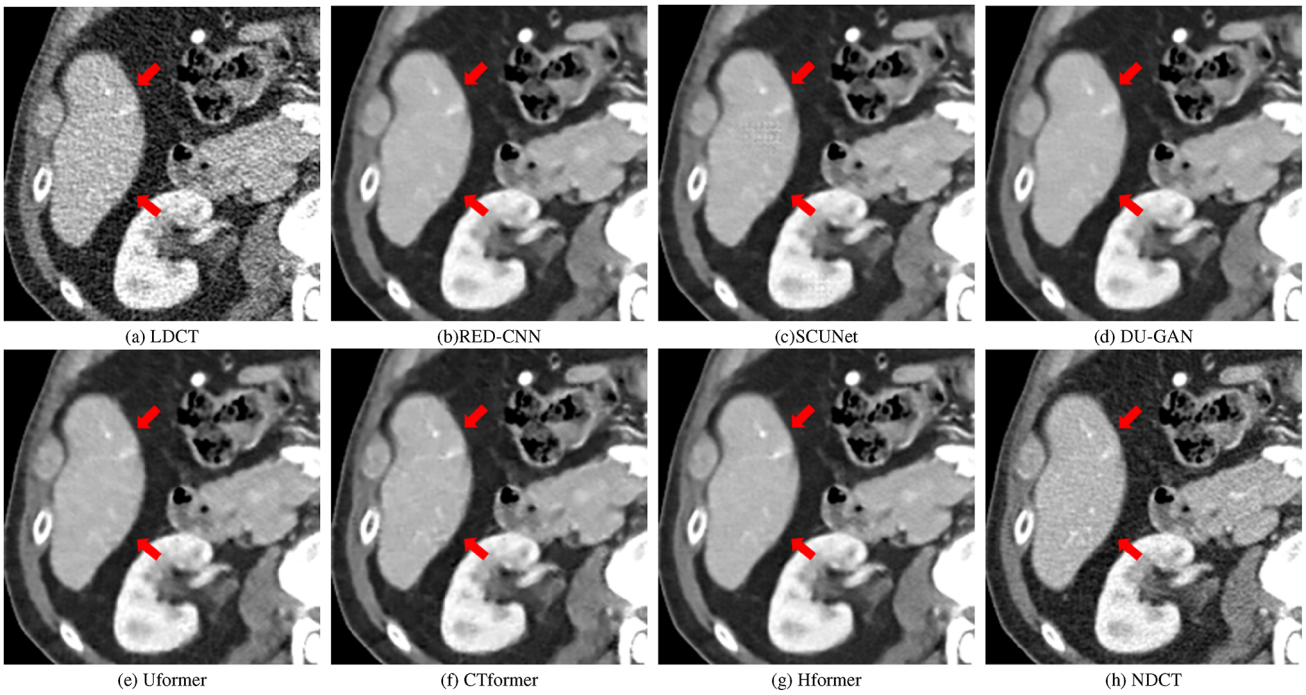

Fig. 7 The corresponding ROI of Fig. 6

and CTformer completely reconstructed the internal vessels of the liver, and the details of Hformer are more clearly depicted. The other networks caused different degrees of smoothing of the textural details of the soft tissues. Although CTformer can also obtain a better tissue structure, it is significantly inferior to Hformer in terms of noise suppression performance. In summary, Hformer can effectively use the advantages of convolution and self-attention to effectively reconstruct the tissue structure while reducing noise and preserving more clinically useful information.

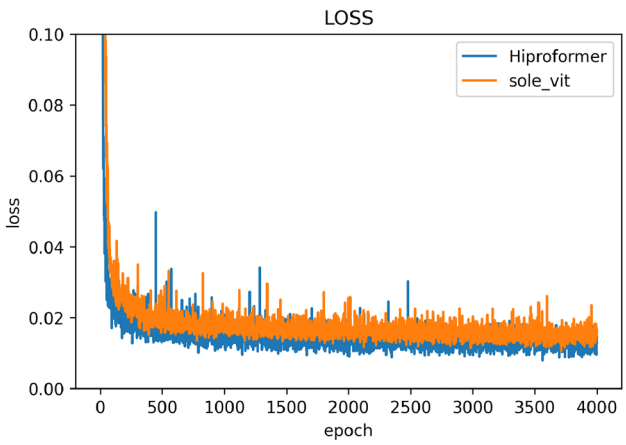

Fig. 8 LOSS visualization of Hformer and Sole-ViT on case L506 after different iterations

4.3 Ablation study

Impact of Hformer blocks. Hformer blocks are used in our network to enhance feature integration during the feature extraction phase. Compared with VIT, which uses only the self-attention mechanism, the Hformer block integrates the inherent advantages of convolution and self-attentiveness in the feature extraction process. To verify the effectiveness of this component, a single ViT model without the Hformer block was designed. We use convolution only in the downsampling stage, with a convolution kernel size of 3×3 and a step size of 2. We subsequently employ a five-layer transform for feature extraction and denoising purposes, utilizing an identical embedding size. The results of the visual comparison are illustrated in Fig. 9a–d. Finally, we can clearly see that Sole-ViT brings additional speckle organization by examining the connected area within the marked region in Fig. 9e–h. In addition, Fig. 8 and Table 4 show that Hformer

Table 4 Quantitative evaluation results of the Sole-ViT, the Hformer, and the Hformer with different numbers of blocks

NET	Block	#param. (M)	RMSE	SSIM	PSNR
Hformer	1	1.65	8.6956	0.9163	33.4405
Sole-ViT	1	1.99	9.2224	0.9089	32.9161
Hformer	2	1.68	8.7677	0.9154	33.3664
Hformer	4	1.75	8.8271	0.9148	33.3046

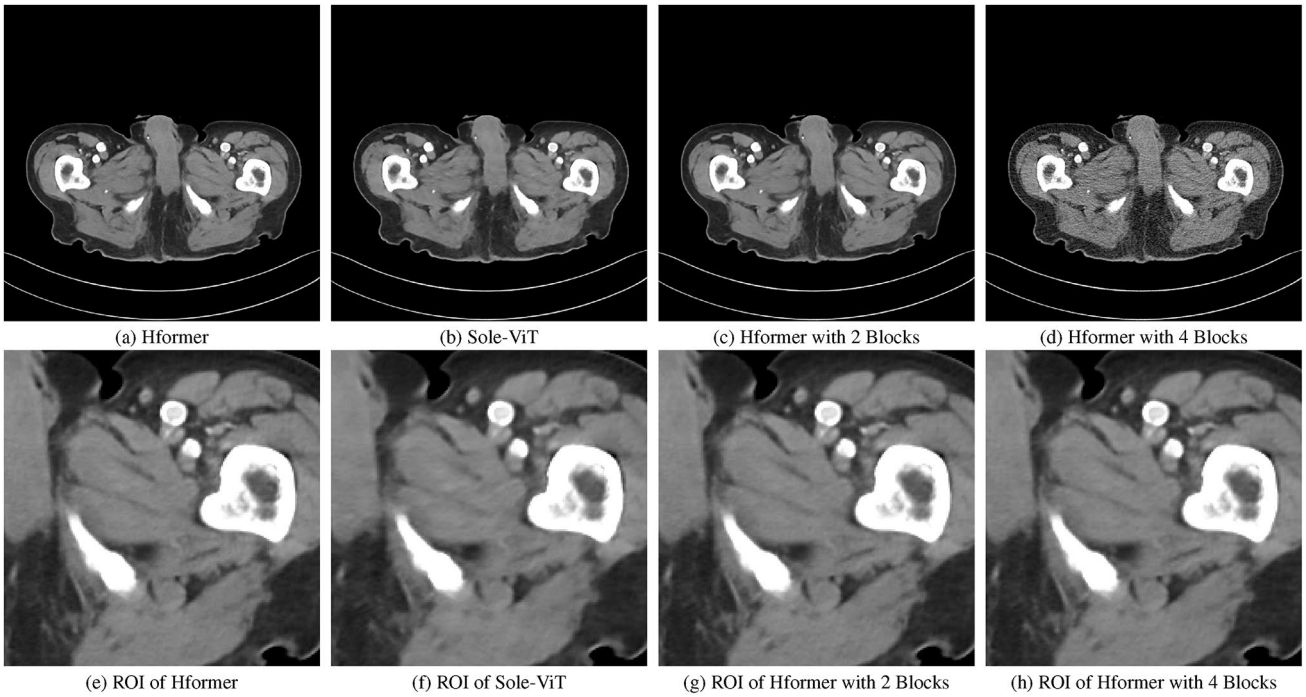

Fig. 9 The performance of Hformer on case L506 with lesion Pelvic Bone. **a** LDCT, **b** Solve-ViT, **c** Hformer with 2 blocks, **d** Hformer with 4 blocks. **e**–**h** are the corresponding magnified ROIs from (**a**–**d**)

converges faster than Sole-Vit with a difference of 0.5244 for PSNR, 0.0074 for SSIM, and 0.5268 for RMSE.

Impact of Hformer numbers. We investigated the impact on the network performance by adjusting the number of Hformer modules in Fig. 1. The number of modules was set to 1, 2, and 4 blocks. As the number of data blocks increases, the depth of the network increases and the computational cost also increases slightly. Table 4 shows that only one Hformer module yields better performance than the Hformer with more blocks.

5 Conclusion

In this study, we designed a novel fast LDCT denoising model. The core of the model is referred to as the Hformer, which combines the advantages of both CNN and local self-attention. We used the well-known dataset AAPM-Mayo Clinic Low-Dose CT Grand Challenge Dataset to evaluate and validate the performance of our proposed Hformer and compare it with the latest SOTA method. The simulation results show that our model achieves excellent results in terms of noise suppression and structural protection, with an effective reduction in the number of training parameters.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Shi-Yu Zhang and Zhao-Xuan Wang, who contributed equally. The first draft of the manuscript was written by Shi-Yu Zhang, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Statement The data that support the findings of this study are openly available in Science Data Bank at https://www.doi.org/10.57760/sciencedb.j00186.00063 and http://resolve.pid21.cn/31253.11.sciencedb.j00186.00063.

References

1. Y. Yang, W. Fang, X. Huang et al., Static superconducting gantry-based proton CT combined with X-ray CT as prior image for FLASH proton therapy. Nucl. Sci. Tech. **34**(1), 11 (2023). https://doi.org/10.1007/s41365-022-01163-2
2. D. Brenner, E. Hall, Computed tomography-an increasing source of radiation exposure. New Engl. J. Med. **357**, 2277–2284 (2007). https://doi.org/10.1056/NEJMra072149

3. J. Jing, W. Xia, M. Hou et al., Training low dose CT denoising network without high quality reference data. Phy. Med. Bio. **67**, 84002 (2022). https://doi.org/10.1088/1361-6560/ac5f70
4. K. He, X. Zhang, S. Ren et al., *Deep residual learning for image recognition*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
5. F. Fan, D. Wang, H. Guo et al., On a sparse shortcut topology of artificial neural networks. IEEE Trans. Artif. Intell. **3**, 595–608 (2021). https://doi.org/10.1109/TAI.2021.3128132
6. X. Wang, R. Girshick, A. Gupta et al., *Non-local neural networks*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7794–7803
7. A. Vaswani, N. Shazeer, N. Parmar et al., *Attention is all you need*, in *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA* (2017). https://doi.org/10.48550/arXiv.1706.03762
8. Z. Liu, Y. Lin, Y. Cao et al., *Swin transformer: Hierarchical vision transformer using shifted windows*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022. https://doi.org/10.48550/arXiv.2103.14030
9. L. Yuan, Y. Chen, T. Wang et al., *Tokens-to-token vit: Training vision transformers from scratch on imagenet*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 558–567. https://doi.org/10.48550/arXiv.2101.11986
10. F. Fan, J. Xiong, M. Li et al., On interpretability of artificial neural networks: a survey. IEEE Trans. Radiat. Plasma Medical Sci. **5**, 741–760 (2021). https://doi.org/10.1109/TRPMS.2021.3066428
11. F. Yang, H. Yang, J. Fu, *Learning texture transformer network for image super-resolution*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 5791–5800. https://doi.org/10.48550/arXiv.2006.04139
12. H. Wu, B. Xiao, N. Codella et al., *Cvt: Introducing convolutions to vision transformers*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 22–31. https://doi.org/10.48550/arXiv.2103.15808
13. M. Chen, A. Radford, and R. Child et al., *Generative pretraining from pixels*, in *International Conference on Machine Learning. PMLR* (2020), pp. 1691–1703
14. S. Abnar, W. Zuidema, *Quantifying attention flow in transformers*. arXiv: 1048550/arXiv.2005.00928 (2020). https://doi.org/10.48550/arXiv.2005.00928
15. G. Montavon, A. Binder, S. Lapuschkin et al., *Layer-wise relevance propagation: an overview*, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (2019), pp. 193–209
16. D. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006). https://doi.org/10.1109/TIT.2006.871582
17. E. Sidky, X. Pan, Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Phys. Med. Biol. **53**, 4777–4807 (2013). https://doi.org/10.1088/0031-9155/53/17/021
18. Y. Chen, D. Gao, N. Cong, Bayesian statistical reconstruction for low-dose x-ray computed tomography using an adaptive-weighting nonlocal prior. Comput. Med. Imag. Graphics **33**, 495–500 (2009). https://doi.org/10.1016/j.compmedimag.2008.12.007
19. Q. Xu, H. Yu, X. Mou, Low-dose X-ray CT reconstruction via dictionary learning. IEEE Trans. Med. Imaging **31**, 1682–1697 (2012). https://doi.org/10.1109/TMI.2012.2195669
20. J. Cai, X. Jia, H. Gao et al., Cine cone beam ct reconstruction using low-rank matrix factorization: Algorithm and a proof-of-principle study. arXiv:1204.3595 (2012). https://doi.org/10.48550/arXiv.1204.3595
21. Z. Li, L. Yu, J. Trzasko et al., Adaptive nonlocal means filtering based on local noise level for ct denoising. Med. Phys. **41**, 011908 (2014). https://doi.org/10.1118/1.4851635
22. M. Aharon, M. Elad, A. Bruckstein et al., K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE T. Signal Proc. **54**, 4311–4322 (2006). https://doi.org/10.1109/TSP.2006.881199
23. Y. Chen, X. Yin, L. Shi et al., Improving abdomen tumor low-dose ct images using a fast dictionary learning based processing. Phys. Med. Biol. **58**, 5803 (2013). https://doi.org/10.1088/0031-9155/58/16/5803
24. P. Feruglio, C. Vinegoni, J. Gros, Block matching 3d random noise filtering for absorption optical projection tomography. Phys. Med. Biol. **55**, 5401–5415 (2010). https://doi.org/10.1088/0031-9155/55/18/009
25. P. Kaur, G. Singh, P. Kaur, A review of denoising medical images using machine learning approaches. Curr. Med. Imaging Rev. **14**, 675–685 (2018). https://doi.org/10.2174/1573405613666170428154156
26. H. Chen, Y. Zhang, W. Zhang et al., Low-dose ct via convolutional neural network. Biomed. Opt. Express **8**, 679–694 (2017). https://doi.org/10.1364/BOE.8.000679
27. T. Würfl, F. Ghesu, V. Christlein et al., *Deep learning computed tomography. in International conference on medical image computing and computer-assisted intervention*, in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016. MICCAI 2016*, ed by S. Ourselin, L. Joskowicz, M. Sabuncu et al. (Springer, 2016), pp. 432–440. https://doi.org/10.1007/978-3-319-46726-9_50
28. H. Chen, Y. Zhang, M. Kalra et al., Low-dose ct with a residual encoder-decoder convolutional neural network. IEEE T. Med. Imaging **36**, 2524–2535 (2017). https://doi.org/10.1109/TMI.2017.2715284
29. K. Chen, L. Zhang, J. Liu et al., Robust restoration of low-dose cerebral perfusion CT images using NCS-Unet. Nucl. Sci. Tech. **33**, 30 (2022). https://doi.org/10.1007/s41365-022-01014-0
30. J. Liu, Y. Zhang, Q. Zhao et al., Deep iterative reconstruction estimation (dire): approximate iterative reconstruction estimation for low dose ct imaging. Phys. Med. Biol. **64**, 135007 (2019). https://doi.org/10.1088/1361-6560/ab18db
31. Y. Ma, Y. Ren, P. Feng et al., Sinogram denoising via attention residual dense convolutional neural network for low-dose computed tomography. Nucl. Sci. Tech. **32**, 41 (2021). https://doi.org/10.1007/s41365-021-00874-2
32. W. Xia, Z. Lu, Y, Huang, et al., CT Reconstruction with PDF: parameter-dependent framework for multiple scanning geometries and dose levels. IEEE Trans. Med. Imaging **40**, 3065–3076 (2021). https://doi.org/10.1109/TMI.2021.3085839
33. Z. Lu, W. Xia, Y. Huang et al., M3NAS: multi-scale and multi-level memory-efficient neural architecture search for low-dose CT denoising. IEEE Trans. Med. Imaging **42**, 850–863 (2022). https://doi.org/10.1109/TMI.2022.3219286
34. L. Huang, H. Jiang, S. Li et al., wo stage residual cnn for texture denoising and structure enhancement on low dose ct image. Comput. Meth. Prog. Biomed. **184**, 105115 (2020). https://doi.org/10.1016/j.cmpb.2019.105115
35. C. Tan, Q. Chao, M. Yang et al., A selective kernel-based cycle-consistent generative adversarial network for unpaired low-dose CT denoising. Precis. Clin. Med. **5**, pbac011 (2022). https://doi.org/10.1093/pcmedi/pbac011
36. Z. Zhang, L. Yu, X. Liang et al., *Transct: dual-path transformer for low dose computed tomography*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2021), pp. 55–64. https://doi.org/10.48550/arXiv.2103.00634
37. A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., *An image is worth 16 ×16 words: Transformers for image recognition at scale*. arXiv: 2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929

38. H. Chen, Y. Wang, T. Guo et al., *Pre-trained image processing transformer*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vol. 12 (2021), pp. 299–310. https://doi.org/10.48550/arXiv.2012.00364
39. M. Kumar, D. Weissenborn, N. Kalchbrenner, Colorization transformer (2021). https://doi.org/10.48550/arXiv.2102.04432
40. Z. Wang, X. Cun, J. Bao et al., *Uformer: a general u-shaped transformer for image restoration*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17683–17693
41. A. Luthra, H. Sulakhe, T. Mittal et al., Eformer: edge enhancement based transformer for medical image denoising. arXiv: 2109.08044 (2021). https://doi.org/10.48550/arXiv.2109.08044
42. D. Wang, F. Fan, Z. Wu et al., Ctformer: convolution-free token-2token dilated vision transformer for low-dose ct denoising. arXiv: 2202.13517 (2022). https://doi.org/10.48550/arXiv.2202.13517
43. B. Graham, A. El-Nouby, H. Touvron et al., *Levit: a vision transformer in convnet's clothing for faster inference*. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 12239–12249. https://doi.org/10.48550/arXiv.2104.01136
44. K. Zhang, Y. Li, J. Liang et al., Practical blind denoising via swin-conv-unet and data synthesis. arXiv: 2203.13278 (2022). https://doi.org/10.48550/arXiv.2203.13278
45. Z. Dai, H. Liu, Q.V. Le et al., Coatnet: marrying convolution and attention for all data sizes. Adv. Neur. Inform. Proc. Syst. **34**, 3965–3977 (2021). https://doi.org/10.48550/arXiv.2106.04803
46. F. Wu, A. Fan, A. Baevski et al., Pay less attention with lightweight and dynamic convolutions. arXiv: 1901.10430 (2019). https://doi.org/10.48550/arXiv.1901.10430
47. Z. Wu, Z. Liu, J. Lin et al., Lite transformer with long-short range attention. arXiv: 2004.11886 (2020). https://doi.org/10.48550/arXiv.2004.11886
48. A. Gulati, J. Qin, C. Chiu et al., Conformer: Convolution-augmented transformer for speech recognition. arXiv: 2005.08100 (2020). https://doi.org/10.48550/arXiv.2005.08100
49. A. Srinivas, T. Lin, N. Parmar et al., *Bottleneck transformers for visual recognition*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 16519–16529. https://doi.org/10.48550/arXiv.2101.11605
50. S. d'Ascoli, H. Touvron, M. L. Leavitt et al., *Convit: improving vision transformers with soft convolutional inductive biases*, in *International Conference on Machine Learning. PMLR* (2021), pp. 2286–2296. https://doi.org/10.48550/arXiv.2107.06263
51. J. Guo, K. Han, H. Wu et al., *Cmt: convolutional neural networks meet vision transformers*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12175–12185. https://doi.org/10.48550/arXiv.2107.06263
52. X. Mao, C. Shen, Y. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections (2016). https://doi.org/10.48550/arXiv.1603.09056
53. K. He, X. Zhang, S. Ren, *Identity mappings in deep residual networks*, in *European Conference on Computer Vision*, vol. 9908 (Springer, Cham, 2016), pp. 630–645
54. J. Xie, L. Xu, E. Chen, *Image denoising and inpainting with deep neural networks*, in *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1 (2012), pp. 341–349
55. Q. Han, Z. Fan, Q. Dai et al., *On the connection between local attention and dynamic depth-wise convolution*, in *International Conference on Learning Representations*. arXiv: 2106.04263 (2021). https://doi.org/10.48550/arXiv.2106.04263
56. J. Ba, J. Kiros, G. Hinton et al., Layer normalization. arXiv: 1607.06450 (2016). https://doi.org/10.48550/arXiv.1607.06450
57. D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus). arXiv: 1606.08415 (2016). https://doi.org/10.48550/arXiv.1606.08415
58. C. McCollough, A. Bartley, R. Carter et al., Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. Med. Phys. **44**, e339–e352 (2017). https://doi.org/10.1002/mp.12345
59. Z. Huang, J. Zhang, Y. Zhang et al., DU-GAN: Generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising. IEEE T. Instrum. Meas. **71**, 1–12 (2021). https://doi.org/10.1109/TIM.2021.3128703