Signal, Image and Video Processing (2024) 18:1215–1226
https://doi.org/10.1007/s11760-023-02809-3

ORIGINAL PAPER

A residual multi-scale feature extraction network with hybrid loss for low-dose computed tomography image denoising

Lina Jia[1] · Aimin Huang[1] · Xu He[1] · Zongyang Li[1] · Jianan Liang[1]

Received: 23 February 2023 / Revised: 7 September 2023 / Accepted: 25 September 2023 / Published online: 1 November 2023

Abstract
In order to suppress noise and artifacts in low-dose computed tomography (LDCT), various deep learning techniques, especially encoder-decoder networks, have been introduced to improve the quality of LDCT images. However, in the encoder-decoder convolutional neural network, fixed-size convolution kernel, continuous down-sampling operation, and the mean square error (MSE) objective function are used, which cause problems such as low utilization of image information, image information loss, and over-smoothing of denoised image. To improve the quality of reconstructed CT images, in this paper, a LDCT image denoising network based on residual multi-scale feature extraction and hybrid loss function is proposed. On the one hand, the multi-scale feature extraction module is designed and introduced into the residual connection to improve the utilization of image feature information; on the other hand, zero padding is used to solve the information loss problem caused by continuous down-sampling operations, and batch normalization (BN) layer is used to alleviate the over-fitting problem caused by network deepening. In addition, a hybrid loss function consisting of MSE loss, structural similarity (SSIM) loss, and perceptual loss is introduced to generate denoised images with high relevance to human perception. Experimental results show that the proposed algorithm can not only improve the quality of denoised images, but also greatly improve the computational speed compared with the state of the art algorithms.

Keywords Image denoising · Low-dose CT · Convolutional neural networks · Residual multi-scale feature extraction · Hybrid loss

1 Introduction

Due to its fast imaging speed and non-invasive, X-ray computed tomography (CT) plays an important role in modern medical diagnosis [1]. Meanwhile, the radiation exposure and radiation hazard have gained considerable attention [2]. Low-dose computed tomography (LDCT) technology is one of the most common ways to reduce radiation hazard. However, with the decrease of X-ray radiation dose, serious artifacts and noise appear in reconstructed CT images, which directly caused difficulties in disease diagnosis [3].

Over the past few decades, researchers have proposed a number of excellent post-processing methods to improve the LDCT image quality, such as non-local mean method [4, 5] and block matching (BM3D) method [6]. However, due to the complexity of noise distribution, the effectiveness of these algorithms is limited.

The rapid development of deep learning provides a new research direction for image processing. Chen et al. proposed a LDCT residual encoder-decoder convolutional neural network (RED-CNN) [7], which combines a deconvolutional network [8–10] and fast connection [11–13] into a convolutional neural network (CNN) to improve the network. RED-CNN achieves good results both quantitative indicators and subjective vision; however, the denoising results are over-smooth. Yang et al. proposed a LDCT image denoising algorithm based on generative adversarial network (GAN) [14]. The algorithm utilizes both Wasserstein distance and perceptual similarity to reduce noise while preserving information. Li et al. proposed a residual attention module (RAM). By inserting the module RAM into RED-CNN and WGAN separately, the two network RED-CNN-RAM and WGAN-RAM were constructed [15]. Although obtained good results in LDCT image denoising task, the model is complex and time-consuming. Liang et al.[16] proposed a new trainable

✉ Lina Jia
jln_ty@163.com

[1] School of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China

Sobel convolution and used it to design an edge enhancement module, which preserved the details well. Wang et al. [17] proposed a convolution-free Token2Token dilated vision transformer (CTformer) for LDCT denoising, which effectively eliminates the common boundary artifacts.

In addition, the selection of the objective function also directly affects the quality of the denoised image. One of the most popular loss function is the MSE loss function, which calculates the squared average of the pixel-by-pixel error between the denoised image and the normal dose CT (NDCT) image. Although high PSNR values are obtained by using it, the image is inevitably over-smoothed. To address this problem, researchers have investigated many loss function for LDCT image denoising, such as perceptual loss, adversarial loss, and edge loss. In natural image denoising task, researchers have developed SSIM loss, multi-scale structural similarity (MS-SSIM) loss [21], contrast regularization loss [22], etc., which can retain the details of the denoised image well.

Inspired by the aforementioned studies, we proposed an encoder-decoder-based LDCT image denoising network MSFREDCNN. The main contributions of our work are summarized as follows:

We proposed a lightweight multi-scale feature extraction (MSFE) module, which enhance information utilization of the input image.

1. We included zero padding to ensure that the input and output images have the same size, which can reduce the loss of structural information in the input image caused by continuous down-sampling. In addition, a BN layer is implemented in the denoising task to reduce overfitting caused by network deepening.
2. We developed a weighted hybrid loss function consisting of MSE loss, perceptual loss, and SSIM loss to guide the network training. It improves the over-smoothing phenomenon in the denoised images effectively.

The remainder of this paper is organized as follows. Section 2 introduces the theory related to LDCT image denoising and the network architecture proposed in this paper. Section 3 analyzes the experimental results in detail. Section 4 discusses the experimental results in depth and presents a summary of the paper.

2 Theory

This section introduces the theory related to LDCT image denoising and the network architecture proposed in this paper. Section 2.1 explains the denoising model from a mathematical point of view. Section 2.2 outlines the recommended network architecture. Section 2.3 describes the proposed objective function and its mathematical expression in this paper.

2.1 Denoising model

In general, let $X \in R^{H\times W}$ be an LDCT image and $Y \in R^{H\times W}$ be its corresponding NDCT image. Mathematically speaking, the relationship between X and Y can be expressed as follows:

$$X = F(Y) \tag{1}$$

where $F(\cdot)$ represents the complex degradation process from NDCT images to LDCT images.

The core of the denoising problem is to find an operator $T(\cdot)$ such that:

$$\arg\min_{T} \|T(X) - Y\|_2^2 \tag{2}$$

where $T(\cdot) = F^{-1}(\cdot)$.

2.2 The proposed network architecture

This section describes the proposed network architecture in detail. Section 2.2.1 introduces the operational procedures of down-sampling and up-sampling. Section 2.2.2 details the proposed multi-scale feature extraction module. Section 2.2.3 describes the overall network structure and parameters of the proposed LDCT image denoising algorithm in detail.

2.2.1 Down-sampling and up-sampling operations

This algorithm is implemented using the classical encoder-decoder network architecture. In the traditional encoder-decoder network architecture, the encoder is composed of consecutive down-sampling modules,zero-padding and the decoder is composed of consecutive up-sampling modules. With the continuous down-sampling, the output image becomes smaller and smaller. It will inevitably cause the loss of image details, degrading the final denoised image qualification. Different from the rules contained in the traditional codec network structure, in our proposed model, a zero padding operation is used to ensure the same size of input image and output image. By this operation, it can effectively reduce the loss of image information.

Figure 1a shows the operation process of down-sampling in the traditional encoder-decoder network, and Fig. 1b shows the down-sampling operation process used in this paper. Let the input image size be $H \times W$, from Fig. 1a, it can be seen that after the down-sampling operation, the output image size is $(H-2)\times(W-2)$, and the reduction of the image size

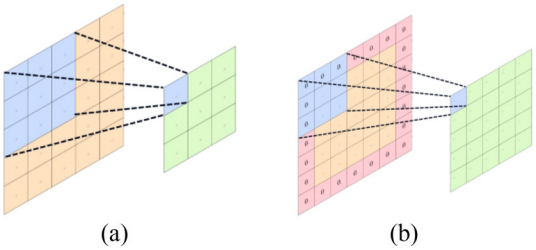

Fig. 1 Down-sampling operation process with stride 1. **a** Operation process of previous down-sampling. **b** Zero-padding sampling operation process

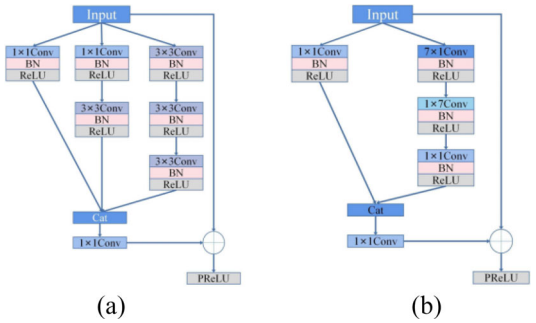

Fig. 2 The proposed multi-scale feature extraction module. **a** MSFEA Module, **b** MSFEB Module

will cause the loss of detail information. After successive down-sampling, the information loss is more obvious. From Fig. 1b, it can be seen that after the down-sampling operation process proposed in this algorithm, the output image size and the input image remain the same, thus reducing the loss of detail information.

2.2.2 Residual multi-scale network structure

In traditional CNN denoising models, a large proportion of the models extract image features by convolutional kernels of fixed size, which often results in poor information utilization of the input image and leads to poor denoising performance. There are also some models that improve their denoising performance by simply stacking the number of convolutional layers. It is found that as the number of convolutional layers increases, the model has more parameters and is more likely to lead to gradient disappearance or network overfitting. Inception network architecture can alleviate both problems to a large extent.

Inspired by Inception_ResNet [23], in this paper, we proposed two MSFE modules to improve the image information utilization. The MSFEA module is shown in Fig. 2a, which divides the input into four paths for multi-scale feature extraction. The four paths consist of three convolutional branches and one directly connected branch, respectively. The convolution kernel sizes of the three convolutional branches are (1 $\times$ 1), (1 $\times$ 1, 3 $\times$ 3), and (3 $\times$ 3, 3 $\times$ 3, 3 $\times$ 3), respectively. While the number of channels are set to 16, (16, 16), and (16, 24, 32), respectively. Following, the output results of the three convolutional branches are concatenated by dimension and passed through a convolutional layer with a convolutional kernel size of 1 $\times$ 1 and a channel number of 96. Finally, the output results are element-wise summed with the directly connected branches. The result of MSFEA module is fed into the PReLU layer of the main denoising network. The proposed MSFEB module is shown in Fig. 2b, which divides the input into three paths for multi-scale feature extraction. The three paths consist of two convolutional branches and one directly connected branch, respectively. The convolution kernel sizes of the 2 convolution branches are (1 $\times$ 1) and (7 $\times$ 1, 1 $\times$ 7, 1 $\times$ 1), respectively. The number of channels are set to 96 and (80, 64, 96), respectively. Then, the output results of the 2 convolutional branches are concatenated by dimension and passed through a convolutional layer with a convolutional kernel size of 1 $\times$ 1 and a channel number of 96. Following, the output of the two convolutional branches are element-wise summed with the directly connected branches. Finally, the result of MSFEB module is fed into the PReLU layer of the main denoising network. In addition, a BN layer is added after each convolutional layer of the MSFE module and activated by the ReLU nonlinear layer to prevent the network from overfitting.

2.2.3 The proposed denoising model

The proposed residual multi-scale feature extraction module is introduced into the classical codec network to form the proposed denoising model MSFREDCNN, as shown in Fig. 3. The model consists of 3 parts: the encoder part, the residual multi-scale feature extraction part, and the decoder part. Unlike RED-CNN, the encoder part of this model consists of 6 convolutional layers, including 2 layers of 5 $\times$ 5 shallow feature extraction layers and 4 layers of 3 $\times$ 3 deep feature extraction layers. Correspondingly, the decoder part also consists of 6 layers of deconvolution layers, which contains 4 layers of 3 $\times$ 3 deconvolution layers and 2 layers of 5 $\times$ 5 deconvolution layers; all the convolution layers use zero padding to ensure the consistency of input and output image sizes, and a BN layer is added after each convolution layer, which can not only speed up the convergence speed and prevent the gradient disappearance or gradient explosion caused by the deepening of the network layers, but also alleviate the problem caused by the overfitting caused by overly complex models or insufficient data sets. The encoder part is activated by the ReLU function, and the decoder part is activated by the PReLU function, which allows the negative part of gradient to be updated with the training of the network as well. In this algorithm, the number of convolutional kernel channels is set to 96 for all except the MSFEA and MSFEB modules.

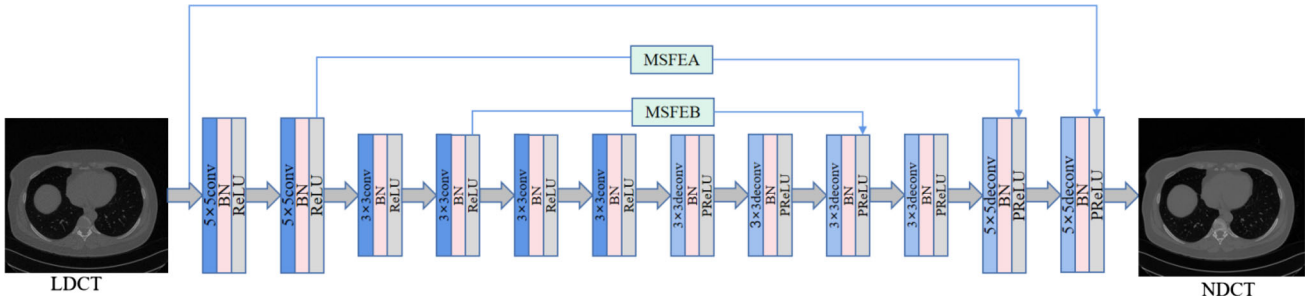

Fig. 3 Overall architecture of our proposed MSFREDCNN model. MSFEA and MSFEB are our proposed multi-scale feature extraction model which is shown in Fig. 2

2.3 Loss function

In addition to the network structure, loss function is an important factor affecting the network denoising performance. In this algorithm, the loss function is designed in two parts: one part is the same as the traditional CNN denoising model, using the MSE loss function to guide the proposed denoising network, highlighting the superiority of the proposed model; the other part is to overcome the shortcomings of the MSE loss function and generate denoised images with high relevance to human perception, alleviating the visual embarrassment of radiologists. In this paper, the weighted sum of the three hybrid objective functions: MSE, SSIM, and perceptual loss, is utilized to guide the proposed denoising model.

2.3.1 MSE loss function

The MSE is the most commonly used loss function in regression models. In the context of LDCT image denoising, it is the mean of the sum of the squared pixel differences between the denoised images and NDCT images. Its formula is shown in Eq. (3).

$$L_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|T(X) - Y\|_2^2 \tag{3}$$

where $T(\cdot)$ denotes the designed denoising network, X denotes the LDCT image, Y denotes the NDCT image, and N denotes the number of pixels in the image$X$.

2.3.2 Perceptual loss

Perceptual loss [24] was first applied in the image style migration task, where it compares the features obtained from the convolution of the generated image with those obtained from the convolution of the real image, making their content and global contextual structure similar. In the context of LDCT image denoising, it represents comparing the features obtained from the convolution of denoised images with those obtained from the convolution of NDCT images, making their content and global structure similar, and guiding the denoising network to generate images that are increasingly close to NDCT. The process can be described by Eq. (4).

$$L_{\mathrm{per}}(X_{\mathrm{denoised}}, Y) = \sum_{i=1}^{M} \|\Phi_i(X_{\mathrm{denoised}}) - \Phi_i(Y)\|^2 \tag{4}$$

where X_{denoised} denotes the denoised image generated by the denoising model, $\Phi_i(\cdot)$ denotes the i-th layer feature map extracted from the feature extraction network, and M denotes the number of layers. The feature extraction network used in this paper is the pretrained model VGG19.

2.3.3 SSIM loss function

It is well known that SSIM is a common metric used to measure the similarity of two images. SSIM is considered to be correlated with the quality perception of the human visual system (HVS). A higher SSIM indicates that the two images are more similar. It can be expressed as follows:

$$\mathrm{SSIM}(X_{\mathrm{denoised}}, Y) = \frac{(2\mu_{X_{\mathrm{denoised}}}\mu_Y + c_1)(2\sigma_{X_{\mathrm{denoised}}Y} + c_2)}{(\mu^2_{X_{\mathrm{denoised}}} + \mu^2_Y + c_1)(\sigma^2_{X_{\mathrm{denoised}}} + \sigma^2_Y + c_2)} \tag{5}$$

where $\mu_{X_{\mathrm{denoised}}}$, μ_Y denotes the mean of and respectively, $\sigma_{X_{\mathrm{denoised}}Y}$ denotes the covariance of X_{denoised} and Y. $\sigma^2_{X_{denoised}}$ and σ^2_Y denote the variance of X_{denoised} and Y respectively. c_1, c_2 are constants to maintain stability.

Based on Eq. (5), the SSIM loss function can be expressed as:

$$L_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}(X_{\mathrm{denoised}}, Y) \tag{6}$$

Finally, we use the weighted sum of several loss functions above to guide the whole denoising network. The total loss

function can be expressed as follows:

$$L_{\text{total}} = \lambda_1 L_{\text{MSE}} + \lambda_2 L_{\text{per}} + \lambda_3 L_{\text{SSIM}} \tag{7}$$

where λ_1,λ_2 and λ_3 are weighting factors.

3 Experimental results and analysis

In this section, we first describe the dataset and the experimental setup details. Next, we elaborate on the evaluation indicators used in this study. Finally, we compared the performance of the proposed MSFREDCNN with five different state-of-the-art denoising methods (RED-CNN-RAM [15], WGAN-RAM [15], CTformer [17], EDCNN [16], and RED-CNN [7]) and visualized the LDCT image denoising results to verify the effectiveness of the proposed network.

3.1 Datasets and experimental setup

Dataset. In the experiments, we used the Low-dose CT image and projection data from Mayo clinic [25]. All CT scans were acquired at routine dose levels for the practice at which they were obtained using standard-clinical protocols for the anatomical region of interest. Each clinical case was processed to include a second projection dataset at a simulated lower dose level. Head and abdomen cases are provided at 25% of the routine dose, and chest cases are provided at 10% of the routine dose. The slice numbers of chest, abdomen, and head data set in our experiments are 5310, 2630, and 1595, respectively. The size of image is 512 × 512. The percentages of images used for training and testing are 70% and 30%, respectively. In our experiments, a fivefold cross-validation was used to train and test the proposed network.

Experimental setup. In the training process, the Adam optimizer [26] was used. Patch_size was set to 64 × 64, the sliding interval is 10, and the experimental batch size was set to the maximum of a single GPU memory. The initial learning rate was set to 1e−4, which is reduced by half every 2000 iterations; the total number of epochs was set to 200. All settings in the comparison algorithm are consistent with the original paper. For a fair comparison, all experiments were trained in the Pytorch 1.11 environment and accelerated using the NIVID RTX3080-10 GB.

3.2 Evaluation of indicators

In this paper, PSNR, SSIM, and root-mean-square error (RMSE) are used to quantitatively evaluate the experimental results: since the SSIM has been described in detail in subsection 2.3.3., it is not repeated here.

PSNR is the ratio of the maximum power of a signal to the noise power that may affect its representation accuracy. Generally speaking, the higher the PSNR value, the better the quality of the generated denoised image. Mathematically, it can be expressed as follows:

$$\begin{aligned}\text{PSNR} &= 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) = 20 \cdot \log_{10}\left(\frac{\text{MAX}}{\sqrt{\text{MSE}}}\right) \\ &= 20 \cdot \log_{10}(\text{MAX}) - 10 \cdot \log_{10}(\text{MSE})\end{aligned} \tag{8}$$

where MAX is the maximum pixel value of the NDCT image.

RMSE is the square root of the square of the deviation of the observed value from the true value and the square root of the ratio of the number of observations K. Mathematically, it can be expressed as:

$$\text{RMSE} = \left[\frac{1}{K}\sum_{i=1}^{K}(X_{\text{denoised}} - Y)^2\right]^{1/2} \tag{9}$$

where K denotes the number of test images.

3.3 Experimental results

3.3.1 Chest dataset

Figures 4 and 5 show the denoised chest CT images obtained by using different denoising models. Obviously, LDCT images have severe noise and artifacts that make diagnosis difficult. RED-CNN-RAM and WGAN-RAM introduce additional noise due to their limited ability in processing images with high level noise. CTformer and EDCNN still retain some noise because they focus much attention on preserving detail in the image generation, mistaking some noise as details. The denoising results of RED-CNN and MSFREDCNN are over-smoothed because single MSE loss are used to average the denoised image, resulting in loss of structural details. The denoising result of the proposed network MSFREDCNN + hybrid loss is the closest to NDCT images. This is mainly because of our proposed residual multi-scale feature extraction model, which makes the network powerful in utilizing contextual information. This is important in dealing with large noisy images.

Table 1 shows the average quantitative performance of results obtained from different denoising methods on the whole chest test set. AVGPSNR is the average peak signal-to-noise ratio over the entire chest test set, AVGSSIM is the average structural similarity over the entire chest test set, and AVGRMSE is the average root-mean-square error over the entire chest test set. The meanings of AVGPSNR, AVGSSIM, and AVGRMSE appeared in the following text are the same as here. The results of RED-CNN-RAM have the lowest average PSNR, SSIM, and RMSE values. The results of WGAN-RAM and CTFormer are relatively closer. The results of our algorithm is the highest. In all, the noise

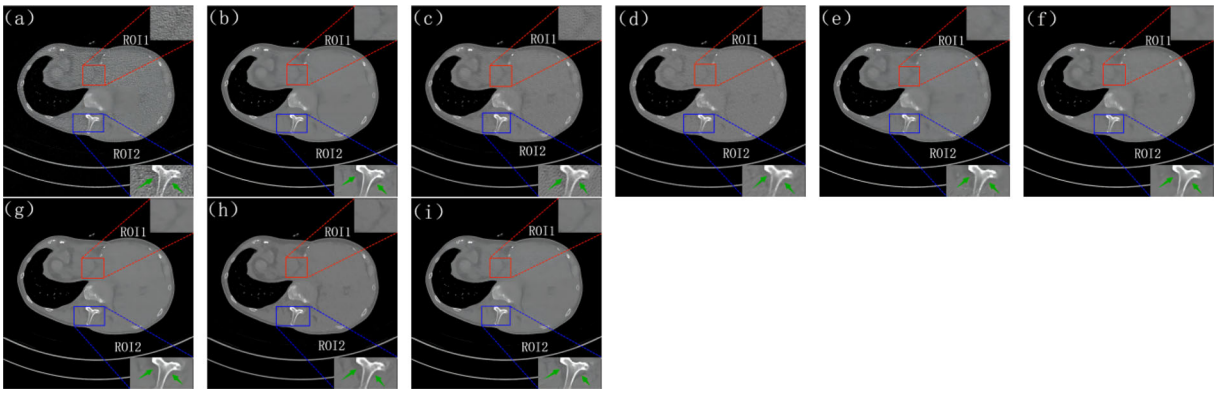

Fig. 4 The denoising results of different models in chest dataset. ROI1 and ROI2 are enlarged in the upper right and lower right corner, respectively. **a** LDCT, **b** NDCT, **c–g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

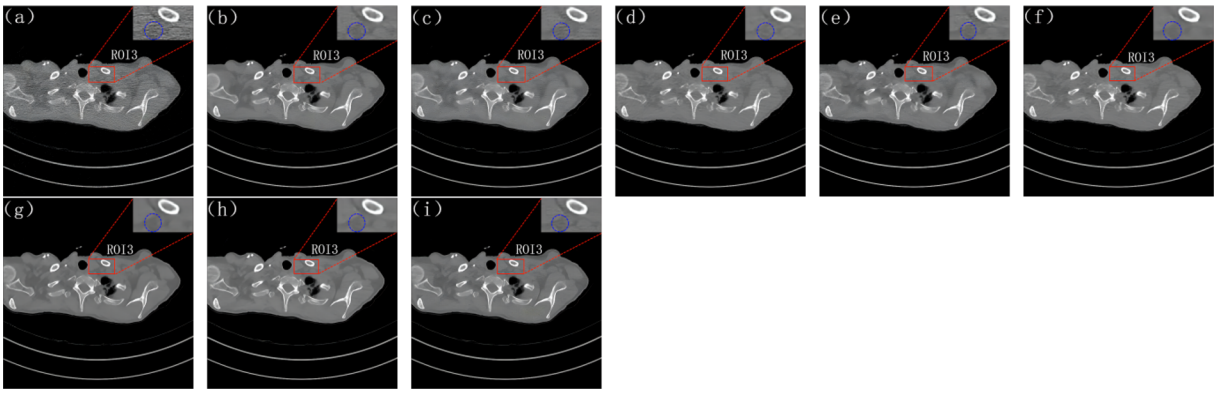

Fig. 5 The denoising results of different models in chest dataset. ROI3 is enlarged in the upper right corner. **a** LDCT, **b** NDCT, **c–g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

and artifact suppression degree after denoising by different methods is in the following order:

RED-CNN-RAM < WGAN-RAM < CTFormer < Hybrid Loss (OUR) < EDCNN < RED-CNN < MSFREDCNN (OUR). According to the run times in Table 1, it can be seen the efficiency of different denoising methods is as follows: CTFormer < RED-CNN-RAM < WGAN-RAM < Hybrid Loss (OUR) < EDCNN < MSFREDCNN (OUR) < RED-CNN. Excepted for RED-CNN, the efficiency of our algorithm is improved highly compared with other algorithms. Figure 6 shows the quantitative evaluation of ROIs in Figs. 4 and 5. These metrics are obtained by averaging ROIs at the same locations on different slices in the whole chest dataset. Obviously, compared with other state-of-the-art algorithms, our model has achieved the highest PSNR, the highest SSIM, and the lowest RMSE.

3.3.2 Abdomen dataset

Figures 7 and 8 show the denoised abdomen CT images obtained from different denoising methods. All of the listed denoising methods can suppress the noise and improve the visual effect to some extent. However, in the region pointed by the blue arrows labeled by ROI4, and the blue circles labeled by ROI5, there is over-smoothing in the results of RED-CNN, which is due to that it only uses the MSE loss as the objective function. Although WGAN-RAM greatly improves visual fidelity, its traditional classification discriminator only provides global structural feedback to the generator due to the use of adversarial loss, and some residual artifacts can still be observed. Comparing the ROI4 and ROI5 in Figs. 7 and 8, we can see that the visual effect of CTformer and EDCNN are better than WGAN-RAM, but there are still residual artifacts in the denoised results. In Figs. 7h–i and 8h–i, the proposed method outperforms other methods in suppressing noise and retaining details.

Table 1 The average quantitative performance of results obtained by using different models on the whole chest test set

Methods	AVGPSNR	AVGSSIM	AVGRMSE	Times [/S]
LDCT	26.7559	0.7415	0.0473	0
RED-CNN-RAM	32.8750	0.8723	0.0227	65,436.6
WGAN-RAM	33.5403	0.8887	0.0211	66,506.8
CTformer	33.5521	0.8799	0.0205	61,200.3
EDCNN	35.1385	0.9088	0.0176	28,955.4
RED-CNN	35.4624	0.9043	0.0169	**8410**
MSFREDCNN (OUR)	**36.2403**	**0.9224**	**0.0156**	25,284.7
Hybrid Loss (OUR)	34.2548	0.9002	0.0189	33,469.9

The best results under each indicator are highlighted in bold font

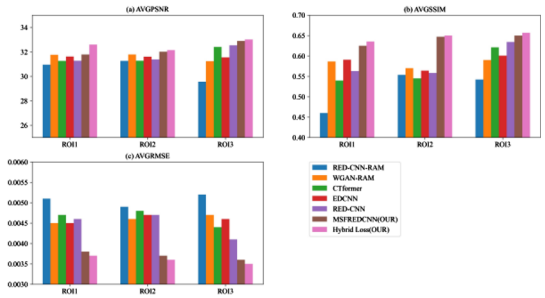

Fig. 6 The average quantitative evaluation of ROIs in Figs. 4 and 5

Table 2 shows the average quantitative performance of results obtained from different denoising networks on the whole abdomen test set. The results of WGAN-RAM are the lowest. The results of RED-CNN-RAM are better than those of WGAN-RAM. The results of CTformer, EDCNN, and RED-CNN are relatively closer. The results of the proposed MSFREDCNN is the highest. Excepted for RED-CNN, the efficiency of our algorithm is improved highly compared with other all algorithms. Figure 9 shows the quantitative evaluation of ROI in Figs. 7 and 8. Obviously, no matter in ROI4 or ROI5, the proposed network has achieved the highest PSNR, the highest SSIM, and the lowest RMSE.

3.3.3 Head dataset

In order to further verify the generalization performance of the proposed algorithm, experiments on head dataset are conducted in this section. The visual comparison is shown in Figs. 10 and 11. The quantitative evaluation is shown in Table 3. The results are consistent with the above chest and abdomen dataset, that is, the proposed algorithm is best in both visual comparison and quantitative indicators. In addition, Fig. 12 presents the changes of the PSNR with the increase of epochs during training head dataset. It is not difficult to find that our model has faster convergence and higher PSNR value.

3.4 Hyper-parameter experiments

In this section, the hyper-parameters of different loss functions are analyzed. In this study, we set λ_1 to 1 based on the

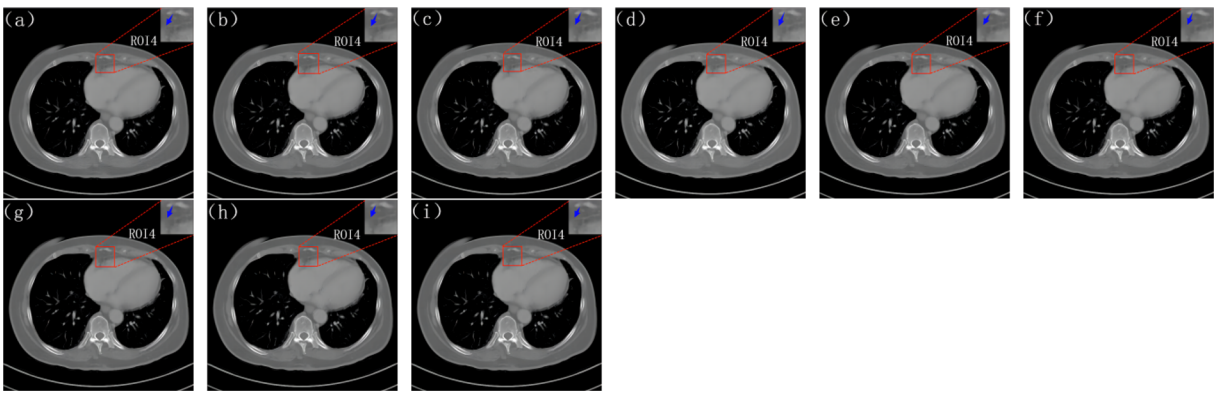

Fig. 7 The denoising results of different models in abdomen dataset. ROI4 is enlarged in the upper right corner. **a** LDCT, **b** NDCT, **c**–**g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

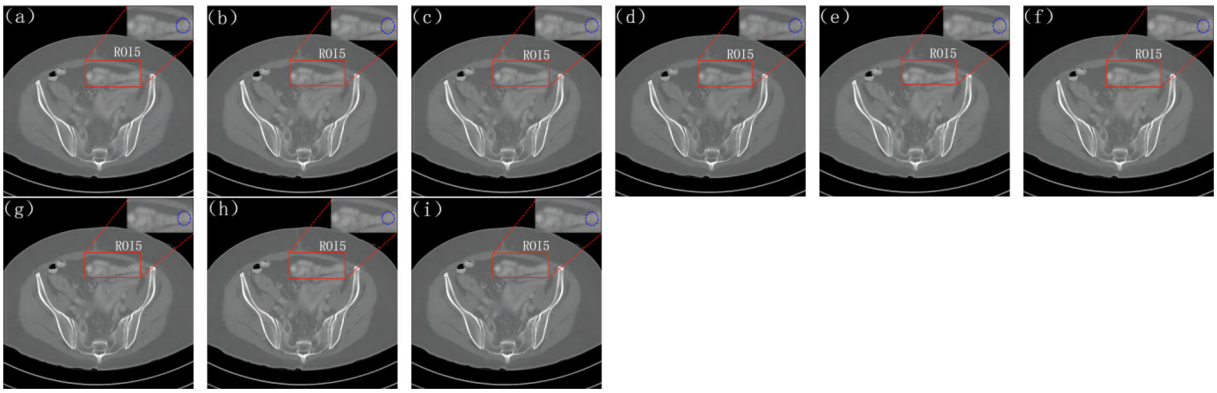

Fig. 8 The denoising results of different models in abdominal dataset. ROI5 is enlarged in the upper right corner. **a** LDCT, **b** NDCT, **c**–**g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

Table 2 The average quantitative performance of results obtained by using different models on the whole abdomen test set

Methods	AVGPSNR	AVGSSIM	AVGRMSE	Times [/S]
LDCT	40.3027	0.9264	0.0098	0
RED-CNN-RAM	42.2910	0.9678	0.0076	35,391.1
WGAN-RAM	41.3149	0.9533	0.0087	30,893.1
CTformer	43.4655	0.9684	0.0069	27,925.4
EDCNN	43.8361	0.9703	0.0067	9987.9
RED-CNN	43.9573	0.9720	0.0064	**5093.1**
MSFREDCNN (OUR)	**44.3343**	**0.9733**	**0.0061**	9985.0
Hybrid Loss (OUR)	44.2327	0.9727	0.0062	15,538.5

The best results under each indicator are highlighted in bold font

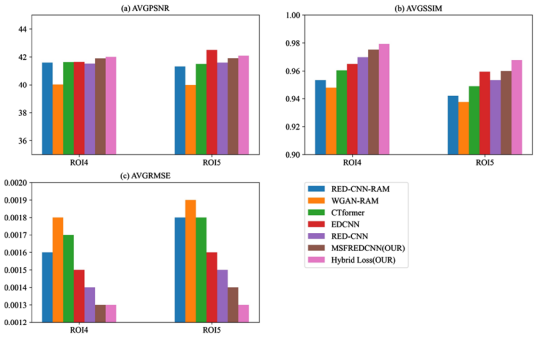

Fig. 9 The average quantitative evaluation of ROIs in Figs. 7 and 8

importance of MSE in the overall objective function. The values of λ_2 and λ_3 were determined by parameter selection experiments. In particular, by analyzing the performance of the average PSNR values obtained by using proposed MSFREDCNN with different $\lambda_2 : \lambda_3$ in three datasets, the optimum values of λ_2 and λ_3 were determined. Table 4 shows the average PSNR values of MSFREDCNN under different $\lambda_2 : \lambda_3$. Obviously, when $\lambda_2 : \lambda_3 = 2 : 1$, the average PSNR values are the highest (as shown in bold in the Table 4). Therefore, was set in this study. In addition, Table 5 shows the influence of different λ_2 and λ_3 values on the average PSNR values in different datasets when $\lambda_2 : \lambda_3 = 2 : 1$. It is not difficult to find that the average PSNR values of the model are the highest when $\lambda_2 = 0.02$ and $\lambda_3 = 0.01$. Therefore, $\lambda_2 = 0.02$ and $\lambda_3 = 0.01$ were set in this study.

3.5 Ablation experiments

In this section, four ablation experiments are performed to prove the effectiveness of each module in the proposed algorithm: The network structure, the objective function, the patch size, and the sliding interval. For the convenience of description, we define the following symbolic representations.

BL: Baseline model. It represents a rough network without any enhancement modules.

BL + PA + BN: It represents the network that adds BN layers and zero padding to the BL model.

BL + PA + BN + MSFE: It represents the network that adds the proposed residual multi-scale feature extraction modules MSFEA and MSFEB to the BL + PA + BN model.

MSE: Using the MSE loss function to guide the proposed MSFREDCNN network.

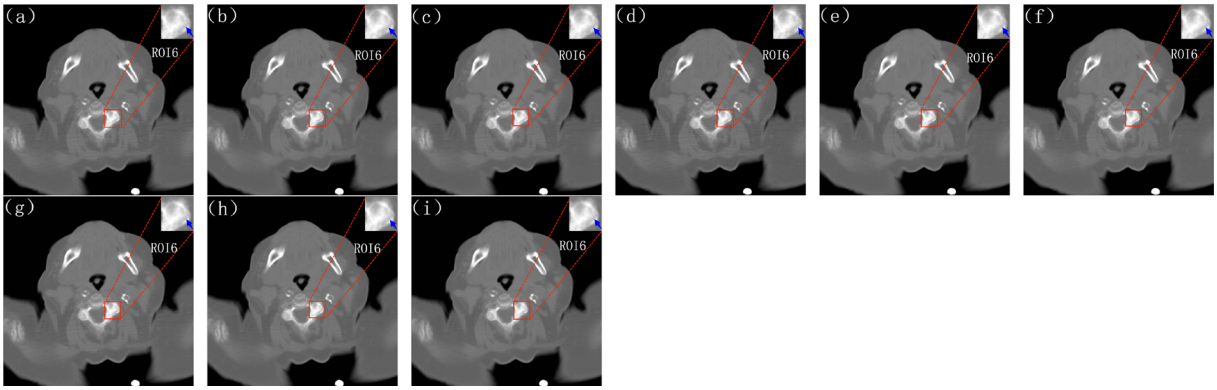

Fig. 10 The denoising results of different models in head dataset. ROI6 is enlarged in the upper right corner. **a** LDCT, **b** NDCT, **c–g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

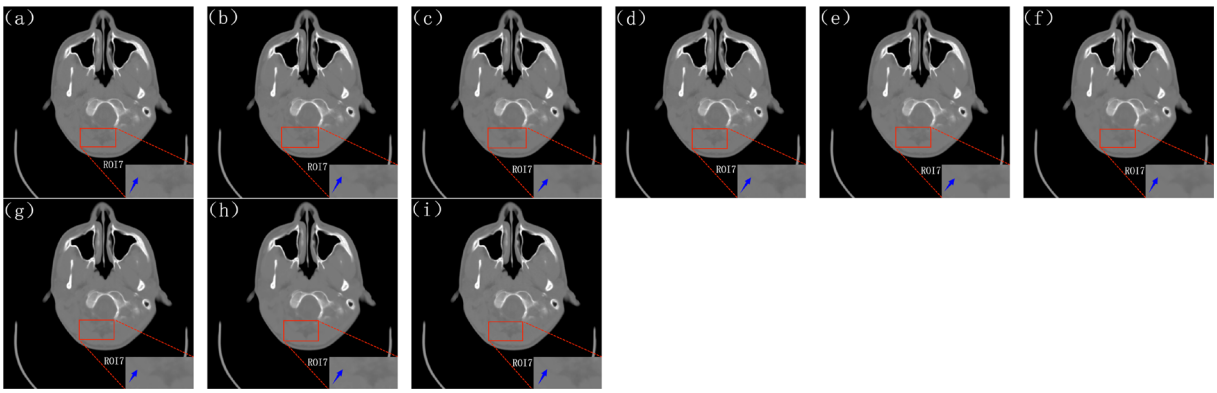

Fig. 11 The denoising results of different models in head dataset. ROI7 is enlarged in the upper right corner. **a** LDCT, **b** NDCT, **c–g** are the denoised results of RED-CNN-RAM, WGAN-RAM, CTformer, EDCNN and RED-CNN, **h** MSFREDCNN + MSE Loss (OUR), **i** MSFREDCNN + Hybrid Loss (OUR)

Table 3 The average quantitative performance of results obtained by using different models on the whole head test set

Methods	AVGPSNR	AVGSSIM	AVGRMSE	Times [/S]
LDCT	44.1609	0.8912	0.0079	0
RED-CNN-RAM	46.4744	0.9162	0.0074	51,016.5
WGAN-RAM	45.5045	0.9004	0.0077	49,551.4
CTformer	47.4550	0.9279	0.0063	17,097.3
EDCNN	47.3641	0.9235	0.0063	7138.3
RED-CNN	47.2316	0.9222	0.0065	**3803.7**
MSFREDCNN (OUR)	**47.5572**	**0.9401**	**0.0059**	6997.1
Hybrid Loss (OUR)	47.5381	0.9386	**0.0059**	13,433.9

The best results under each indicator are highlighted in bold font

MSE + VGG: Using MSE and perceptual loss function to guide the proposed MSFREDCNN network.

MSE + SSIM: Using MSE and SSIM loss function to guide the proposed MSFREDCNN network.

MSE + SSIM + VGG: Using the proposed hybrid loss function consisting of MSE, SSIM, and perceptual loss function to guide the proposed MSFREDCNN network.

For the network structure, the modules to be ablated are added to the baseline model each time. The objective results of the ablation experiments are shown in Table 6. It can be observed that the denoising performance of the network gradually improves with the addition of each module.

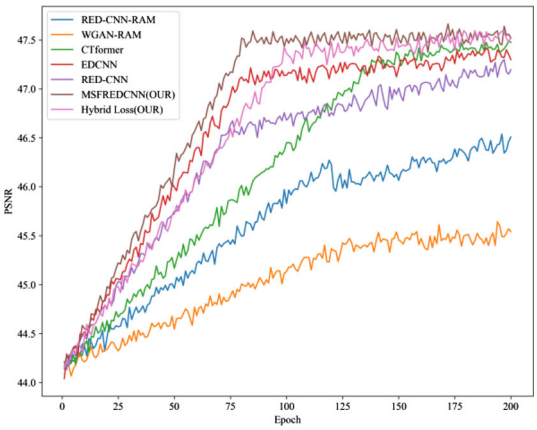

Fig. 12 The changes in average PSNR of different models during training head dataset

Table 4 The average PSNR of MSFREDCNN under different $\lambda_2 : \lambda_3$ on the whole test set

$\lambda_2 : \lambda_3$	AVGPSNR (chest dataset)	AVGPSNR (abdomen dataset)	AVGPSNR (head dataset)
1: 1	34.6729	43.5011	46.7966
1: 2	34.9816	43.6743	46.5201
1: 5	35.0127	43.8820	45.9033
2: 1	**36.2144**	**44.3019**	**47.3441**
5: 1	35.7648	44.0175	46.8015

The best results under each indicator are highlighted in bold font

Table 5 The influence of different λ_2 and λ_3 values on the average PSNR values when $\lambda_2 : \lambda_3 = 2 : 1$

$\lambda_2 : \lambda_3$	AVGPSNR (chest dataset)	AVGPSNR (abdomen dataset)	AVGPSNR (head dataset)
0.002, 0.001	35.6258	43.7535	47.1194
0.02, 0.01	**36.2403**	**44.3344**	**47.5572**
0.2, 0.1	36.1560	44.0013	46.6339
2, 1	34.7709	43.5058	45.8842

The best results under each indicator are highlighted in bold font

Ablation study of the objective function is performed to verify the effectiveness of the proposed hybrid loss function. Four different loss functions are added to the proposed denoising model. Since the background region in CT images occupies a large part of the CT image, which is often not useful for the physician's diagnosis. Therefore, the ROIs mentioned in Sect. 3.3 are used for average quantitative performance calculation. The comparison results are shown in Table 7. It can be seen that the highest objective metrics can be obtained by using the hybrid loss function on the proposed network model.

In addition, it is important to analyze the impact of patch size and sliding interval in the training process. In the training progress, three different patch sizes of 64×64, 128×128, and 256×256 were used for experiments respectively. Table 8 demonstrates the ablation study of patch size on the three test datasets. From Table 8, it appears that smaller image blocks yield superior performance. This is because small patches can capture the local features and details of the image more efficiently, and small patches are more adaptable to different image types. For the parameter sliding interval, we trained the model with sliding intervals of 5, 10, 15, and 20, respectively. Table 9 shows the corresponding experimental results. As shown in Table 9, although the model achieves superior outcomes at small sliding intervals, the computation and processing duration increase correspondingly. Considering a reasonable trade-off between the achieved performance and the involved calculation complexity a sliding interval of 10 was used in our algorithm.

4 Discussion and conclusion

By evaluating the effectiveness of each component in the network in the previous section, it can be concluded that the reasons why the proposed method can achieve good results can be attributed to the following four aspects. Firstly, a lightweight residual multi-scale feature extraction module MSFE is proposed in this paper, which can extract multi-scale features of images from multiple paths and improve the feature utilization of images. Secondly, a zero padding operation is used in the convolution layer to ensure the input and output images are of the same size, which solves the problem of image structure information loss due to continuous down-sampling. Thirdly, due to the over-fitting phenomenon that may be caused by the deepening of the network, we add a batch normalization layer to the network to alleviate this problem. Finally, the hybrid loss function, which consist of MSE loss, perceptual loss, and SSIM loss, helps to improve the performance of the proposed network further. This is because the hybrid loss function synthesizes the characteristics of a single loss function, evaluates and guides the network training from multiple perspectives, making the performance of the network further improved [27].

Our proposed multi-scale feature extraction module is effective to extract rich information from LDCT images. The proposed hybrid loss is beneficial to focus on multiple features of the image and improves the comprehensive utilization rate of the information and features. Finally, the quality of denoised image is improved. The contribution of this paper includes two aspects. On the one hand, we proposed a method to improve the utilization rate of image features and finally improved the denoised image quality effectively. On the other hand, our method provides ideas

Table 6 Network structural ablation experiments

	Chest dataset			Abdominal dataset			Head dataset		
	PSNR	SSIM	RMSE	SSIM	PSNR	RMSE	PSNR	SSIM	RMSE
BL	35.3591	0.9135	0.0169	44.0998	0.9214	0.0064	47.2501	0.9353	0.0062
BL + PA + BN	35.6441	0.9159	0.0163	44.1336	0.9287	0.0063	47.3267	0.9397	0.0062
BL + PA + BN + MSFE	**36.1665**	**0.9225**	**0.0156**	**44.3057**	**0.9389**	**0.0061**	**47.5279**	**0.9400**	**0.0060**

The best results under each indicator for each dateset are highlighted in bold font

Table 7 Objective function ablation experiments

	Chest dataset			Abdominal dataset			Head dataset		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
LDCT	29.0042	0.3198	0.0096	37.7464	0.9165	0.0018	40.4332	0.9322	0.0017
MSE	31.6648	0.6321	0.0042	40.9418	0.9549	0.0013	43.1135	0.9649	0.0014
MSE + VGG	31.7801	0.6329	0.0041	41.1517	0.9582	0.0012	43.5474	0.9677	0.0013
MSE + SSIM	32.7187	0.6951	0.0036	41.7237	0.9611	0.0012	43.6829	0.9710	0.0011
MSE + VGG + SSIM	**32.9558**	**0.7050**	**0.0034**	**42.0383**	**0.9639**	**0.0011**	**44.0057**	**0.9753**	**0.0011**

The best results under each indicator for each dateset are highlighted in bold font

Table 8 The ablation study of patch size on three test datasets

Patch size	Chest dataset			Abdominal dataset			Head dataset		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
64 × 64	**36.2403**	**0.9224**	**0.0156**	**44.3343**	**0.9733**	**0.0061**	**47.5572**	**0.9401**	**0.0059**
128 × 128	35.8415	0.9006	0.0162	44.0879	0.9510	0.0069	46.3217	0.8921	0.0062
256 × 256	35.2573	0.8857	0.0169	42.8249	0.9164	0.0081	45.6312	0.8176	0.0075

The best results under each indicator for each dateset are highlighted in bold font

to improve the utilization rate of image and noise features from more angles and aspects to further improve the quality of LDCT images. Despite MSFREDCNN has achieved satisfactory experimental results, there are still some problems to be solved. First, the noise prior information is not utilized in our model. Second, our network is lack of self-adaptive attention on the most important parts of the image. In the future, we will analyze the characteristics of speckle noise and streak artifacts in LDCT images, study their distinguishing features, and design different noise suppression modules. In addition, we will introduce attention mechanisms, design high-level statistical attention module, and improve the noise reduction ability of the network.

In conclusion, we proposed residual multi-scale feature extraction network with hybrid loss for LDCT image denoising. In our experiments, we analyze the denoising performance of the proposed MSFREDCNN both subjectively and objectively. In addition, ablation experiments for network structure, objective loss, patch size, and sliding interval are performed. Experimental results showed that our method outperforms some related denoising networks.

Table 9 The ablation study of sliding interval on three test datasets

Sliding interval	Chest dataset		Abdomen dataset		Head dataset	
	PSNR	Time [/s]	PSNR	Time [/s]	PSNR	Time [/s]
5	36.4339	33,527.1	44.5210	14,668.3	47.5834	7864.5
10	36.2403	25,284.7	44.3343	9985.0	47.5572	6997.1
15	35.7440	22,364.9	43.5677	7488.4	46.2383	6025.6
20	34.2589	19,677.2	43.1059	6733.2	45.9837	5331.8

Authors' contributions AMH develop the idea and accomplished the manuscript writing and performed the experiments. All authors accomplished the manuscript revising. All authors read and approved the final manuscript.

Funding This work was supported in part by the Natural Science Foundation of Shanxi Province under Grant 202203021222015, in part by the State Council and the central government guide local funds of China under Grant YDZX20201400001547, in part by the postgraduate practice and innovation program of Shanxi Province under Grant 2023SJ011.

Data availability The data and material can be made available on request.

Code availability The code can be made available on request.

Declarations

Conflict of interest We declare that we have no conflict of interest.

References

1. Luo, L., Hu, Y., Chen, Y.: Research status and prospect for low-dose ct imaging. J. Data Acquisition Process. **30**(1), 224–234 (2015)
2. Aliasharzadeh, A., et al.: A survey of computed tomography dose index and dose length product level in usual computed tomography protocol. J. Cancer Res. Ther. **14**(3), 549 (2018)
3. Brenner, D.J., Hall, E.J.: Computed tomography—an increasing source of radiation exposure. N. Engl. J. Med. **357**, 2277–2284 (2007)
4. Zhang, Y., Salehjahromi, Z., Yu, H.: Tensor decomposition and non-local means based spectral CT image denoising. J. J. X-ray Sci. Technol. **27**(3), 397–416 (2012)
5. Luo, Z., Yin, Y., Bi, S.: Denoising algorithm for CT oral image based on Bayesian threshold and Non-local mean. In: C. Computers and Software Engineering (AEMCSE), pp. 288–292. IEEE (2020)
6. Yahya, A., Tan, J., Su, B., et al.: BM3D image denoising algorithm based on an adaptive filtering. J. Multimedia Tools Appl. **79**, 20391–20427 (2020)
7. Chen, H., Zhang, Y., Kalra, M.K., et al.: Low-dose ct with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
8. Dong, J., Roth, S., Schiele, B.: Deep wiener deconvolution: Wiener meets deep learning for image deblurring. Adv. Neural. Inf. Process. Syst. **33**(12), 1048–1059 (2022)
9. Shin, S.-Y., Kim, D.-M., Suh, J.-W.: Image denoiser using convolutional neural network with deconvolution and modified residual network. IEICE Trans. Inf. Syst. **102**(8), 1598–1601 (2019)
10. You, H., Yu, L., Tian, S., et al.: Mc-net: Multiple max-pooling integration module and cross multi-scale deconvolution network. Knowl.-Based Syst. **231**, 107456 (2021)
11. Liu, H., Jin, X., Liu, L.: Low-dose CT image denoising based on improved DD-net and local filtered mechanism. Comput. Intell. Neurosci. (2022)
12. Yang, L., Shangguan, H., Zhang, X., et al.: High-frequency sensitive generative adversarial network for low-dose ct image denoising. IEEE Access. **8**, 930–943 (2019)
13. Rassil, A., Chougrad, H., Zouaki, H.: Augmented graph neural network with hierarchical global-based residual connections. Neural Netw. **150**, 149–166 (2022)
14. Yang, Q., Yan, P., Zhang, Y., et al.: Low-dose ct image denoising a generative adversarial network with wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
15. Li, M., Du, Q., Duan, L., et al.: Incorporation of residual attention modules into two neural networks for low-dose ct denoising. Med. Phys. **48**(6), 2973–2990 (2021)
16. Liang, T., Jin, Y., Li, Y., et al.: EDCNN: edge enhancement-based densely connected network with compound loss for low-dose CT denoising. In: IEEE International Conference on Signal Processing (ICSP), pp. 193–198 (2020)
17. Wang,D., Fan, F., Wu, Z., et al.: CTformer: convolution-free Token2Token dilated vision transformer for low-dose CT denoising. arXiv e-prints (2022)
18. Ma, Y., Wei, B., Feng, P., et al.: Low-dose ct image denoising using a generative adversarial network with a hybrid loss function for noise learning. IEEE Access. **8**, 65719–65729 (2020)
19. Kyung, S., Won, J., Pak, S., et al.: Mtd-Gan: Multi-task discriminator based generative adversarial networks for low-dose CT denoising. In: International Workshop on Medicine Learning for Medical Image Reconstruction, pp 133–144. Springer (2022)
20. Huang, Z., Zhang, J., Zhang, Y., et al.: Du-gan: generative adversarial networks with dual-domain u-net-based discriminators for low-dose ct denoising. IEEE Trans. Instrum. Meas. **71**, 1–12 (2021)
21. Naseem, R., Alaya Cheikh, F., Beghdadi, A., et al.: Cross-modal guidance assisted hierarchical learning based siamese network for mr image denoising. Electronics **10**(22), 2855 (2021)
22. Wu, H., Qu, Y., Lin, S., et al.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10551–10560 (2021)
23. Tang, Y., Gong, W., Chen, X., et al.: Deep inception-residual laplacian pyramid networks for accurate single-image super-resolution. IEEE Trans. Neural Netw. Learn. Syst. **31**(5), 1514–1528 (2019)
24. Han, M., Shim, H., Baek, J.: Perceptual CT loss: implementing CT image specific perceptual loss for CNN-based low-dose CT denoiser. IEEE Access. **10**, 62412–62422 (2022)
25. Moen, T.R., Chen, B., Holmes III, D.R., et al.: Low-dose ct image and projection dataset. Med. Phys. **48**(2), 902–911 (2021)
26. Zhang, Z.: Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), IEEE, pp. 1–2 (2018)
27. Shan, H., Zhang, Y., Yang, Q., et al.: 3-d convolutional enccoder-decoder network for low-dose ct via transfer learning from a 2-d trained network. IEEE Trans. Med. Imaging **37**(6), 1522–1534 (2022)