Original Article

# Ultrasound image denoising autoencoder model based on lightweight attention mechanism

**Liuliu Shi[1,2,3], Wentao Di[1,3], Jinlong Liu[4,5,6]**

[1]School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai, China; [2]Key Laboratory of Power Machinery and Engineering of Ministry of Education, Shanghai Jiao Tong University, Shanghai, China; [3]Shanghai Key Laboratory of Multiphase Flow and Heat Transfer in Power Engineering, Shanghai, China; [4]Institute of Pediatric Translational Medicine, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China; [5]Shanghai Engineering Research Center of Virtual Reality of Structural Heart Disease, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China; [6]Shanghai Institute for Pediatric Congenital Heart Disease, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

*Contributions:* (I) Conception and design: L Shi, W Di; (II) Administrative support: L Shi, J Liu; (III) Provision of study materials or patients: J Liu; (IV) Collection and assembly of data: W Di; (V) Data analysis and interpretation: L Shi, W Di; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Liuliu Shi, PhD. School of Energy and Power Engineering, University of Shanghai for Science and Technology, 516 Jungong Rd., Shanghai 200093, China; Key Laboratory of Power Machinery and Engineering of Ministry of Education, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, China; Shanghai Key Laboratory of Multiphase Flow and Heat Transfer in Power Engineering, 516 Jungong Rd., Shanghai 200093, China. Email: shiliuliu@usst.edu.cn.

**Background:** The presence of noise in medical ultrasound images significantly degrades image quality and affects the accuracy of disease diagnosis. The convolutional neural network–denoising autoencoder (CNN-DAE) model extracts feature information by stacking regularly sized kernels. This results in the loss of texture detail, the over-smoothing of the image, and a lack of generalizability for speckle noise.

**Methods:** A lightweight attention denoise-convolutional neural network (LAD-CNN) is proposed in the present study. Two different lightweight attention blocks (i.e., the lightweight channel attention (LCA) block and the lightweight large-kernel attention (LLA) block are concatenated into the downsampling stage and the upsampling stage, respectively. A skip connection is included before the upsampling layer to alleviate the problem of gradient vanishing during backpropagation. The effectiveness of our model was evaluated using both subjective visual effects and objective evaluation metrics.

**Results:** With the highest peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values at all noise levels, the proposed model outperformed the other models. In the test of brachial plexus ultrasound images, the average PSNR of our model was 0.15 higher at low noise levels and 0.33 higher at high noise levels than the suboptimal model. In the test of fetal ultrasound images, the average PSNR of our model was 0.23 higher at low noise levels and 0.20 higher at high noise levels than the suboptimal model. The statistical analysis showed that the p values were less than 0.05, which indicated a statistically significant difference between our model and the other models.

**Conclusions:** The results of this study suggest that the proposed LAD-CNN model is more efficient in denoising and preserving image details than both conventional denoising algorithms and existing deep-learning algorithms.

**Keywords:** Ultrasound image denoising; deep learning; speckle noise; attention mechanism

Submitted Nov 21, 2023. Accepted for publication Mar 08, 2024. Published online Apr 10, 2024.
doi: 10.21037/qims-23-1654
**View this article at:** https://dx.doi.org/10.21037/qims-23-1654

## Introduction

With the development of medical imaging technology, ultrasound imaging has been widely implemented in clinical medical diagnostics due to its non-invasive, low-cost, and real-time advantages, making it one of the most widely used diagnostic tools in medicine today. However, due to their acquisition mechanism, ultrasound images are inherently subject to speckle noise. The signal generated by the ultrasound probe scatters as it transmits through the organism, resulting in the appearance of speckle noise in the medical ultrasound image. Speckle noise is a predominant contributor to poor image quality in medical ultrasound images (1). In addition to affecting the quality of the ultrasound image, speckle noise can also lead to the loss of important details in the image, thus posing a challenge to the physician in making a diagnosis. Therefore, the efficient removal of speckle noise from ultrasound images by means of scientific algorithms is of great practical importance.

Conventional denoising techniques for ultrasound images are primarily classified into the following two categories: (I) frequency domain filtering; and (II) spatial domain filtering. Frequency domain filtering requires transforming the signal from the spatial domain to the frequency domain through Fourier transform, wavelet transform, or other methods. Denoising is performed in the frequency domain, and the resulting denoised image is obtained via inverse transformation. Rodrigues *et al.* (2) combined the S-median threshold wavelet filter with a bilateral filter to effectively remove speckle noise. Vimalaraj *et al.* (3) introduced a method for denoising ultrasound images that combines the dual-tree complex wavelet transform with the possibility of fuzzy C-means. In spatial domain filtering, the denoising task is accomplished through the direct manipulation of the pixels in the image. Examples of these manipulations include the use of a median filter (4), Frost filter (5), and non-local means filter (6). The aforementioned conventional denoising techniques can result in the loss of image details during the denoising process and suffer from extensive time consumption and manual parameterization.

In recent years, advances in deep learning have led to the development of new image denoising techniques (7-9). Liu *et al.* (10) proposed deep convolutional encoder-decoder networks for image denoising, incorporating a skip connection in the convolutional and transposed convolutional layers. This results in the extraction of more detailed features from the bottom layer and thus improves the denoising effect. Zhang *et al.* (11) presented a feed-forward denoising convolutional neural network (DnCNN) that employs deep network architecture and residual learning in image denoising. Zeng *et al.* (12) proposed the residual encoder-decoder with squeeze-and-excitation network (RED-SENet) based on the channel attention mechanism. The channel attention mechanism (13) was initially applied in the fields of image classification (14) and target monitoring (15). It extracts important information in the channel domain, while the large-kernel attention (LKA) mechanism extracts a larger range of information in the spatial domain. The large-kernel attention mechanism has high performance in image super-resolution (16), image classification (17), and image segmentation (18). Gondara *et al.* (19) proposed a convolutional neural network–denoising autoencoder (CNN-DAE) model for medical image denoising. Encoding and decoding stages form the entire network, and denoising is accomplished by learning the mapping from the noisy images to the original images. The CNN-DAE model extracts feature information by stacking regularly sized kernels; however, this leads to a loss of texture detail and the over-smoothing of the image. Additionally, it only takes Gaussian noise into account, which limits its generalizability to speckle noise.

To address the above-mentioned issues, we modified the structure of the CNN-DAE model and proposed a lightweight attention denoise-convolutional neural network (LAD-CNN) model. Two different lightweight attention blocks [i.e., the lightweight channel attention (LCA) block and the lightweight large-kernel attention (LLA) block] were concatenated into the encoding and decoding stages, respectively. Further, skip connections were used to obtain a larger receptive field to preserve detailed features and avoid overfitting. Experiments were carried out on both public and private data sets. The results indicated that the proposed LAD-CNN model is more efficient in denoising and preserving image details than both conventional denoising algorithms and existing deep-learning algorithms.

## Methods

### *Noise model*

Speckle noise in ultrasound images is signal-dependent (20), it can be modeled as:

$$z = u^{\gamma} \times \mathrm{n} + u \quad [1]$$

where n ~ N (0, $\sigma^2$) is the normal distribution with an expectation of zero and a variance of $\sigma^2$; $u$ is the original

 

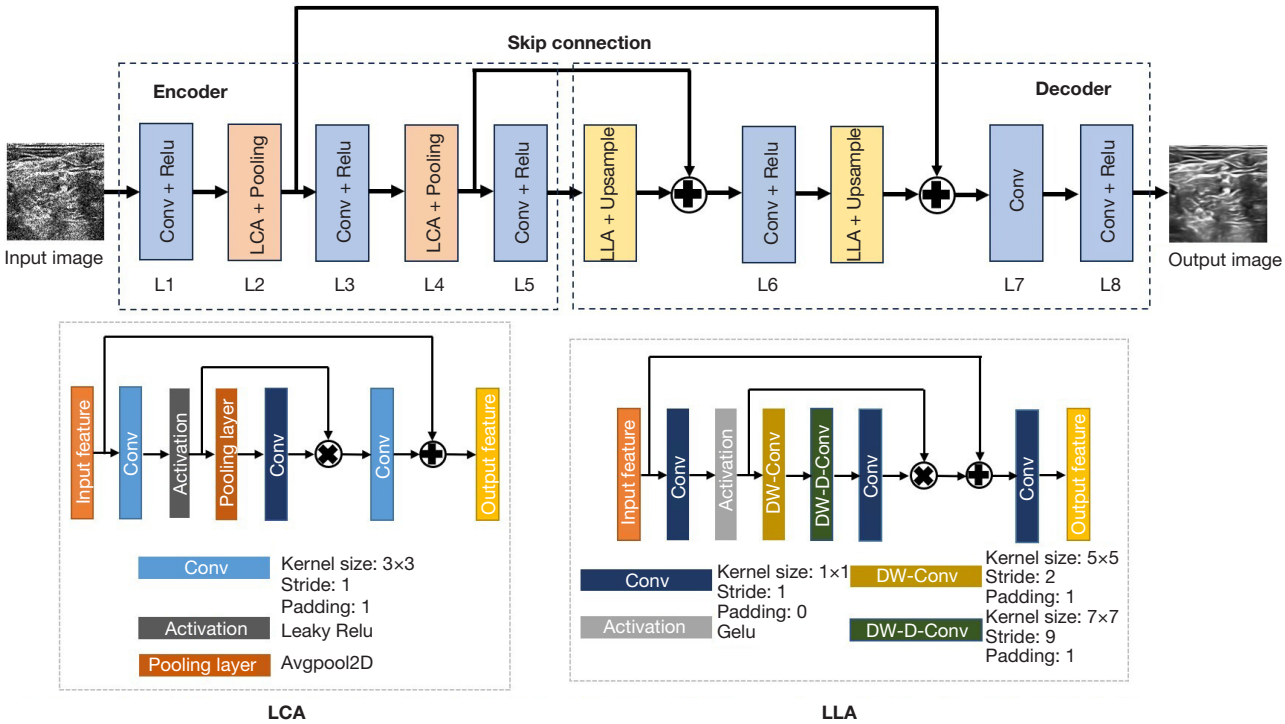


**Figure 1** Structure of the LDA-CNN model. LCA, lightweight channel attention block; LLA, lightweight large-kernel attention block; LAD-CNN, lightweight attention denoise-convolutional neural network; Conv, convolution layers; Relu, rectified linear unit; DW-Conv, depthwise convolution; DW-D-Conv, depthwise dilation convolution; Gelu, Gaussian Error Linear Unit.

image; and $\gamma$ is affected by the ultrasound equipment and other operating factors. The experimental results achieved by Loupas *et al.* (21), based on logarithmically compressed images, showed that the simulated speckle noise image at $\gamma$ =0.5 was consistent with the real ultrasound image. Since then, this model has been used in numerous studies to simulate speckle noise in ultrasound images (20,22-24).

### *Network architecture*

To address the inadequacy in the speckle noise denoising performance of the CNN-DAE model, we propose an autoencoder network for denoising ultrasound images that employs the attention mechanism. To enhance the sensitivity of the network to the texture details in medical ultrasound images, two different lightweight attention blocks are introduced into the downsampling stage and the upsampling stage, respectively. In the current network, two downsampling layers and upsampling layers are retained, with the subpixel convolution being chosen for the upsampling stage. Further, to alleviate the problem of gradient vanishing during backpropagation, a skip connection is included before the upsampling layers as shown in *Figure 1*.

During the encoding stage, the initial noisy image undergoes a convolutional layer with a convolution kernel size of 5×5. This process extends the image from a single-channel image to a 64-channel image while preserving the feature size of the image through padding. Maximum pooling is used in the downsampling stage to extract the most distinctive features from the ultrasound image. This enables the model to detect edge and texture information while reducing its complexity and training time, as it does not require any additional training parameters. To avoid the loss of details due to the convolution operation immediately after the downsampling, a LCA block is added after the downsampling layer to help the model to make more effective use of the feature information.

The low-dimensional features are extracted from the decoding stage through a 3×3 convolutional layer. To prevent the loss of significant information in multiple pooling layers, as well as the gradient vanishing during the data training process, a skip connection is used to concatenate feature maps from the downsampling layer

 

**Table 1** LAD-CNN parameters

| Layer number | Layer type | Encoder |
|---|---|---|
| L1 | Conv + Relu | Kernel size: 5×5; padding: 2; stride: 1; Relu |
| L2 | Pooling | Maxpool 2D; Kernel size: 2×2 |
| L3 | Conv + Relu | Kernel size: 3×3; padding: 1; stride: 1; Relu |
| L4 | Pooling | Maxpool 2D; Kernel size: 2×2 |
| L5 | Conv + Relu | Kernel size: 3×3; padding: 1; stride: 1; Relu |
| L6 | Conv + Relu | Kernel size: 3×3; padding: 1; stride: 1; Relu |
| L7 | Conv | Kernel size: 5×5; padding: 2; stride: 1 |
| L8 | Conv + Relu | Kernel size: 1×1; padding: 0; stride: 1; sigmoid |

LAD-CNN, lightweight attention denoise-convolutional neural network; Conv, convolution layers; Relu, rectified linear unit.

**Table 2** Comparison of total model parameters

| Model | Total parameters |
|---|---|
| RED-SENet [12] | 1,851,169 |
| DnCNN [11] | 556,096 |
| CNN-DAE [19] | 298,497 |
| LAD-CNN | 533,155 |

RED-SENet, residual encoder-decoder with squeeze-and-excitation network; DnCNN, denoising convolutional neural network; CNN-DAE, convolutional neural network-denoising autoencoder; LAD-CNN, lightweight attention denoise-convolutional neural network.

with feature maps from the upsampling layer. The subpixel convolution without training parameters is used to recover the image size (25). This upsampling strategy can be used as a substitute for interpolation or transposed convolution, and it has a larger receptive field while reducing computational consumption and training time. The subpixel convolution technique converts depth to space to extract features from a low-resolution image and to rearrange the pixels from multiple channels into a single channel in a high-resolution image. More detailed information can be obtained using skip connection (12). The upsampling layer is succeeded by the LLA block. The image is restored to its initial size after two upsampling layers, and the denoised image is obtained via a 1×1 convolutional layer and a sigmoid activation function. The specific parameters of the present network are shown in *Table 1*. *Table 2* compares the total parameter differences between the LAD-CNN model and other typical deep-learning denoising models. Notably, our model has fewer parameters than most of the other denoising models. Since our model is based on CNN-DAE with two additional modules (the LCA block and LLA block), it has more parameters than the CNN-DAE model.

### *LCA block*

The channel attention mechanism improves the denoising performance, as the weights of the important feature channels are strengthened adaptively, while the unimportant feature channels are suppressed. In the LAD-CNN model, a LCA block is proposed on the basis of the channel attention mechanism. To reduce the complexity of the model, a 1×1 convolutional layer is used instead of the fully connected layer to extract features and enhance the nonlinear capability of the network. By using a LCA block, the number of training parameters is reduced and the important information in the feature channels is emphasized.

### *LLA block*

The large-kernel convolution mechanism captures relations between pixels that are further apart, enabling long-distance dependence by convolution with a larger-sized convolution kernel. The use of a large-kernel convolution mechanism usually leads to a significant increase in the number of parameters and the computational cost. Guo *et al.* (26) introduced the LKA mechanism by decomposing the convolution operation with a kernel size of k × k into the following three components: a depthwise convolution with a kernel size of k/d × k/d, where d is the expansion rate; a depthwise dilation convolution with a kernel size of (2d-1) × (2d-1); and a channel convolution with a kernel size of 1×1. Based on the LKA mechanism, we propose a LLA block. This block also extracts feature information with a kernel size of 1×1. In addition, we incorporate skip connections into the network to introduce shallow features and gradients into the deep network, leading to a significant improvement in convergence speed and the reduction of gradient vanishing. Further, we employ a group convolution technique to further reduce the number of training parameters. During the upsampling stage, there is a loss of information as the low-dimensional feature map is restored to the denoised image. To address this issue,

 

the LLA block is inserted into the upsampling layer, which adaptively allocates weights to the pixels of the feature map and thereby enhances the use of global information.

### *Experimental environment and training data*

To evaluate the effectiveness of the proposed denoising model, experiments were carried out on both simulated speckle noise ultrasound images and real clinical ultrasound images. All the experiments were implemented on the Pytorch framework and accelerated with the NVIDIA GeForce RTX 3060Ti Graphics Processing Unit (GPU).

The Berkeley Segmentation Dataset (BSD400) (27), which comprises 400 images with a size of 180 px × 180 px, was used for training. To augment the image data, the original 400 images were first randomly rotated or translated and then resized to 128 px × 128 px image blocks, resulting in a data set of 2,000 images. Additionally, speckle noise was added to the resulting 2,000 images. Of these, 1,600 images were used for training, and the remaining 400 images were used for validation.

To assess the robustness and generalizability of the model, and confirm its effectiveness in ultrasound image denoising, we chose a variety of public and private data sets from different organ regions for testing, including the Kaggle brachial plexus ultrasound public data set (28), which contains 5,508 images with a size of 580 px × 420 px; the fetal head ultrasound public data set (29), which contains 1,334 images with a size of 800 px × 540 px; and cardiac ultrasound images, which were provided by Shanghai Children's Medical Center and were acquired using the Philips iU 22 ultrasound system with a two-dimensional probe. The imaging parameters were as follows: frame rate: 51 Hz; depth: 15 cm, thermal index of soft tissue: 0.7; and mechanical index: 1.4. The data set comprised 500 images with a size of 790 px × 630 px. From each of the above-mentioned data sets, 100 representative ultrasound images that captured the diversity of the data set well and had clearer textures were extracted and resized to images with a size of 128 px × 128 px. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was carried out with the approval of the local institutional review board of Shanghai Children's Medical Center Affiliated to Shanghai Jiao Tong University School of Medicine, and written informed consent was obtained from the parents of the patients.

### *Quantitative evaluations*

The following three commonly used evaluation metrics were used to quantitatively assess the effectiveness of the ultrasound image denoising: the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and equivalent number of looks (ENL). The PSNR measures the pixel-by-pixel difference between the denoised image and the original image, while the SSIM evaluates the similarity in terms of brightness, contrast, and structure.

To assess the effectiveness of the model in denoising real ultrasound images, denoising was performed on ultrasound data collected from clinical sources (the cardiac ultrasound images provided by Shanghai Children's Medical Center). Due to the unavailability of original images for comparison, metrics such as the PSNR and SSIM were no longer applicable. To overcome this limitation, the ENL was introduced (30,31). The ENL serves as an indicator for evaluating the smoothness of the images in homogeneous areas, providing a metric for assessing denoising efficiency in the absence of original images. The PSNR, SSIM, and ENL are defined as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{\left(u_{\max}\right)^2}{\mathrm{MSE}\left(u,x\right)} \quad [2]$$

$$\mathrm{SSIM} = \frac{\left(2\mu_x\mu_u + c_1\right)\left(2\sigma_{xu} + c_2\right)}{\left(\mu_x^2 + \mu_u^2 + c_1\right)\left(\sigma_x^2 + \sigma_u^2 + c_2\right)} \quad [3]$$

$$\mathrm{ENL} = \frac{\mu_x^2}{\sigma_x^2} \quad [4]$$

where $\mu_x$ and $\mu_u$ are the mean of denoised image $x$ and original image $u$, respectively; $u_{\max}$ is the maximum value of the original image $u$; $\sigma_{xu}$ is the covariance between $x$ and $u$; $\sigma_x$ and $\sigma_u$ are the standard deviations of $x$ and $u$, respectively; and $c_1$ and $c_2$ are constants that are set to 6.5025 and 58.5225 as proposed in (32).

Higher PSNR and SSIM values indicate a greater similarity between the denoised image and the original image, indicating better denoising performance, while a higher ENL value indicates greater accuracy in the denoised image. The PSNR and SSIM were chosen to measure the denoising performance of images that contained artificial noise, while the ENL was chosen to measure the denoising

 

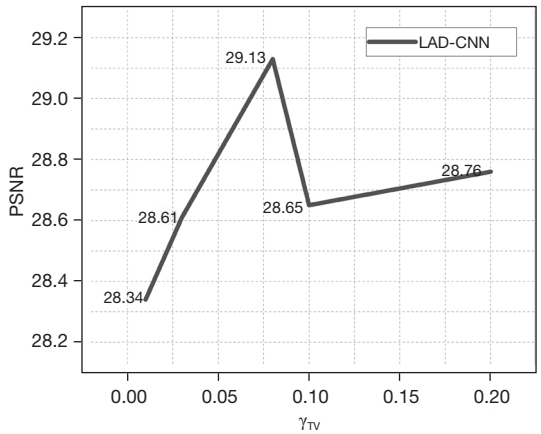


**Figure 2** PSNR curves over the different regularization parameter $\gamma_{TV}$ at σ =3.0. PSNR, peak signal-to-noise ratio; LAD-CNN, lightweight attention denoise-convolutional neural network; $\gamma_{TV}$, predefined weight for the TV loss function; TV, total variation; σ, noise level.

performance of real clinical images that were contaminated by noise (24).

In the training process, the Adam optimizer was implemented with an initial learning rate of 0.001. The learning rate was multiplied by a decay factor of 0.1 if the PSNR of the validation set was monitored and did not increase after five epochs. We trained the model for 120 epochs with a batch size of 10. Weight parameters from the epoch with the highest PSNR and SSIM were chosen for the subsequent tests. To assess the ability of the proposed model to reduce speckle noise, comparative experiments were performed on public and private data sets with conventional denoising methods (block-matching and three-dimensional filtering (BM3D) (33), Gaussian filtering (21), and other deep-learning denoising methods [e.g., DnCNN (11), CNN-DAE (19), and RED-SENet (12)].

### *Loss function*

The mean squared error (MSE) function and the total variation (TV) regularization function are employed to calculate the loss. The MSE function is defined as:

$$L_{MSE} = \frac{\sum_{i=1}^{n}(u-x)^2}{n} \quad [5]$$

where $u$ represents the original image, x represents the denoised image, and n represents the number of the batch size.

The MSE function is effective in training the model; however, artifacts may still occur in the process of image recovery. To address the issue of artifacts, we incorporated the TV regularization function (34) to smooth the image and reduce the artifacts that arise from the denoised image. The TV function is defined as:

$$L_{TV} = \sum_{w=1}^{W}\sum_{h=1}^{H}\sqrt{\left(x^{w+1,h}-x^{w,h}\right)^2+\left(x^{w,h+1}-x^{w,h}\right)^2} \quad [6]$$

where x represents the denoised image, and W and H are the height and width of the image, respectively.

Thus, the overall loss function is defined as:

$$L_{LOSS} = L_{\mathrm{MSE}} + \gamma_{TV} L_{TV} \quad [7]$$

where, $\gamma_{TV}$ is a predefined weight for the TV loss function. To prevent excessive smoothing, $\gamma_{TV}$ should be much smaller than 1 (34). $\gamma_{TV}$ =0.08 was chosen according to the experimental results shown in *Figure 2*. The experiment was carried out at a noise level $\sigma$ =3, and the experimental setup was the same as that for the model training. The test set comprised 100 images taken from the fetal head ultrasound public data set.

## Results

In this study, six noise models with different levels ($\sigma$ =2.0, 3.0, 4.0, 5.0, 6.0, and 7.0) were trained and compared to conventional denoising methods and other deep-learning models. The effectiveness of each model was evaluated based on both subjective visual effects and objective evaluation metrics. *Table 3* presents the average PSNR and SSIM values for each model, under various noise levels, which were based on a test set of the brachial plexus ultrasound images.

As *Table 3* shows, the denoising effects of the deep learning-based methods were significantly better than the conventional denoising models at all noise levels. Among the deep learning-based methods, the current model had the highest PSNR and SSIM value at all noise levels, outperforming the other CNN models, and showed remarkable denoising efficiency compared to the CNN-DAE model.

*Figure 3* shows a subjective visual comparison of the original image, the noised image, and the denoised images. The results showed that BM3D over-smoothed the image, resulting in a significant loss of texture information. Gaussian filtering was ineffective at removing noise from the texture and was inadequate at recovering the edge

 

**Table 3** Comparison of the metrics of the various models using the brachial plexus data set

| Model | Noise level | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | σ =2 | | σ =3 | | σ =4 | | σ =5 | | σ =6 | | σ =7 | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| NOISE | 22.10 | 0.74 | 18.66 | 0.60 | 16.23 | 0.48 | 14.48 | 0.39 | 13.03 | 0.34 | 12.07 | 0.30 |
| BM3D | 26.17 | 0.85 | 23.90 | 0.77 | 22.64 | 0.71 | 20.99 | 0.66 | 20.64 | 0.61 | 19.96 | 0.60 |
| Gauss filter | 23.96 | 0.84 | 22.51 | 0.78 | 21.14 | 0.71 | 18.73 | 0.55 | 18.92 | 0.59 | 18.08 | 0.55 |
| CNN-DAE | 25.37 | 0.89 | 23.88 | 0.85 | 22.83 | 0.82 | 21.99 | 0.78 | 21.23 | 0.75 | 20.56 | 0.72 |
| DnCNN | 26.42 | 0.91 | 24.49 | 0.86 | 22.99 | 0.82 | 22.01 | 0.78 | 21.25 | 0.75 | 20.60 | 0.72 |
| RED-SENet | 22.35 | 0.80 | 23.87 | 0.85 | 21.92 | 0.80 | 20.22 | 0.76 | 20.96 | 0.74 | 20.35 | 0.72 |
| Our model | 26.56 | 0.92 | 24.55 | 0.87 | 23.25 | 0.84 | 22.31 | 0.80 | 21.56 | 0.78 | 20.97 | 0.75 |

σ, noise level; PSNR, peak signal-to-noise ratio; SSIM, structural similarity; NOISE, speckle-noised images; BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

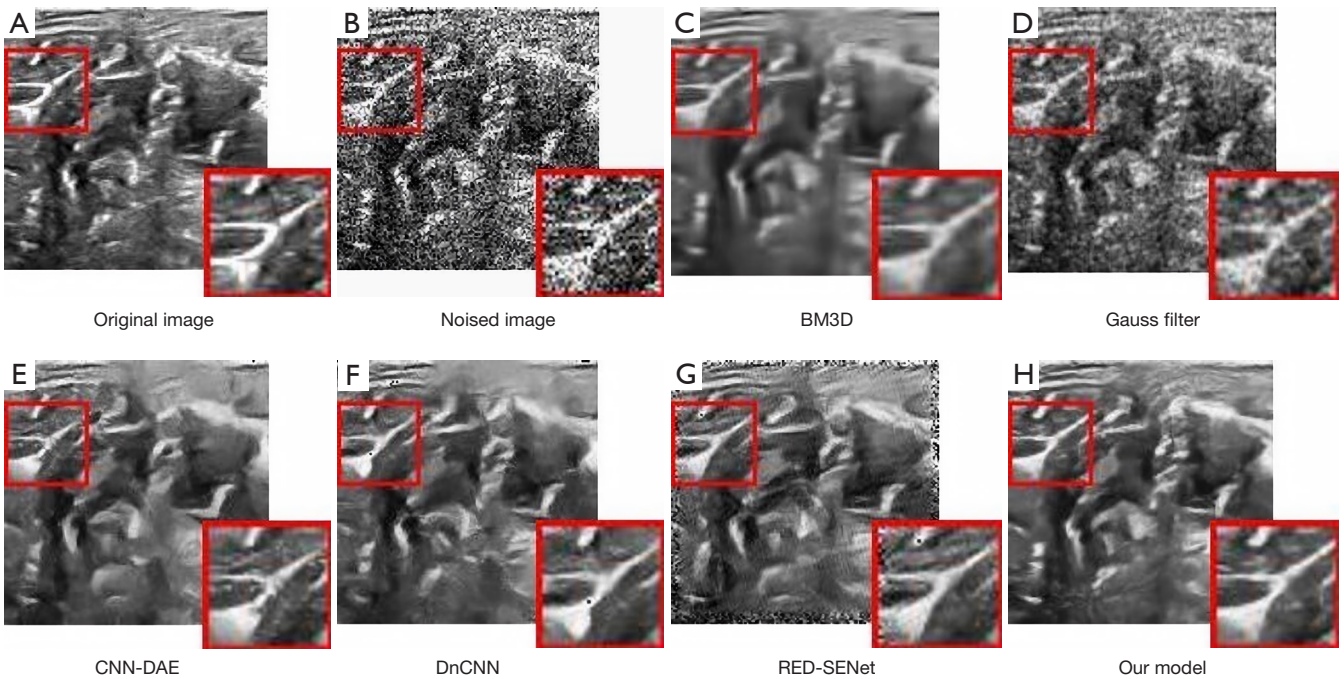


**Figure 3** Subjective visual comparison of the denoised brachial plexus images at σ =5.0. (A) Original image, (B) noised image, (C) BM3D, (D) Gauss filter, (E) CNN-DAE, (F) DnCNN, (G) RED-SENet, (H) our model. BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network; σ, noise level.

 

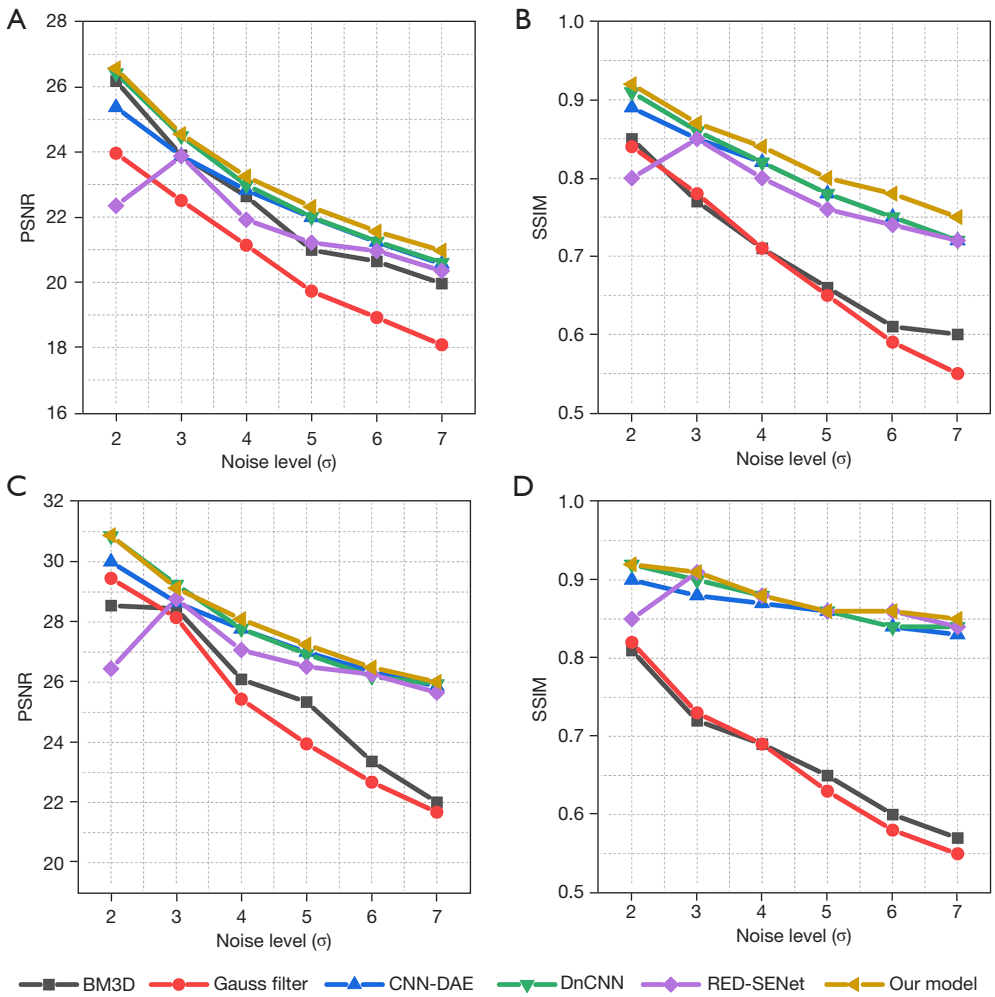


**Figure 4** Variations in the denoising evaluation metrics for each model at different noise levels. (A) The average PSNR for the brachial plexus data set, (B) the average SSIM for the brachial plexus data set, (C) the average PSNR for the fetal head data set, (D) the average SSIM for the fetal head data set. PSNR, peak signal-to-noise ratio; σ, noise level; SSIM, structural similarity; BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

features. The DnCNN model retrieved a greater amount of the texture information and more effectively processed the image edges; however, the denoised image still contained some speckle noise that could subjectively affect visual perception. The RED-SENet model preserved the overall texture information more efficiently, but it faced a significant challenge in the form of edge noise. The LAD-CNN model enabled more effective recognition of edge information than the CNN-DAE model. Additionally, it provided a more comprehensive and detailed representation of texture features. Although the proposed model performed equivalently to the DnCNN in terms of the objective metrics, such as the PSNR and SSIM, on this data set, it was still necessary to evaluate its performance through subjective vision. It is important to note that the image denoised by the DnCNN model had severe white noise. Therefore, from a comprehensive perspective, our model appeared to outperform the DnCNN model.

*Figure 4A*,*4B* illustrate the variation of the evaluation metrics in the denoising performance at different noise levels for each model evaluated on the brachial plexus data set. At low noise levels ($\sigma$ =2, 3, and 4), the average PSNR

**Table 4** Comparison of the metrics of the various models using the fetal head data set

| Model | Noise level | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | σ =2 | | σ =3 | | σ =4 | | σ =5 | | σ =6 | | σ =7 | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| NOISE | 24.58 | 0.68 | 22.93 | 0.59 | 18.81 | 0.47 | 17.10 | 0.40 | 15.82 | 0.35 | 14.80 | 0.32 |
| BM3D | 28.55 | 0.81 | 28.45 | 0.72 | 26.10 | 0.69 | 25.34 | 0.65 | 23.37 | 0.60 | 22.01 | 0.57 |
| Gauss filter | 29.45 | 0.82 | 28.15 | 0.73 | 25.44 | 0.69 | 23.95 | 0.63 | 22.68 | 0.58 | 21.68 | 0.55 |
| CNN-DAE | 30.00 | 0.90 | 28.65 | 0.88 | 27.77 | 0.87 | 27.00 | 0.86 | 26.36 | 0.84 | 25.86 | 0.83 |
| DnCNN | 30.86 | 0.92 | 26.25 | 0.76 | 27.78 | 0.88 | 26.94 | 0.86 | 26.19 | 0.84 | 25.94 | 0.84 |
| RED-SENet | 26.45 | 0.85 | 28.77 | 0.91 | 27.07 | 0.88 | 25.52 | 0.86 | 26.26 | 0.86 | 25.66 | 0.84 |
| Our model | 30.88 | 0.92 | 29.13 | 0.91 | 28.09 | 0.88 | 27.25 | 0.86 | 26.49 | 0.86 | 26.01 | 0.85 |

σ, noise level; PSNR, peak signal-to-noise ratio; SSIM, structural similarity; NOISE, speckle-noised images; BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

of our model was 0.15 higher than that of the suboptimal DnCNN model. At the low noise level of $\sigma$ =2.0, the RED-SENet model could not efficiently identify the speckle noise and image details, ultimately resulting in a lower PSNR and SSIM value than the other convolutional neural network (CNN) models. At high noise levels ($\sigma$ =5, 6, and 7), the average PSNR of our model was 0.33 higher than that of the suboptimal model (DnCNN).

*Table 4* shows the average PSNR and SSIM for each model at different noise levels on the fetal ultrasound images data set. The variations of the denoising performance metrics of each model on the fetal head data set under different noise levels are displayed in *Figure 4C,4D*. At low noise levels ($\sigma$ =2, 3, and 4), the average PSNR of our model was 0.23 higher than that of the suboptimal model. At high noise levels ($\sigma$ =5, 6, and 7), the average PSNR of our model was 0.15 higher than that of the suboptimal model. *Figure 5* shows the subjective visual comparison of the original image, the noised image, and the denoised images. The conventional models had the disadvantages of excessive smoothing and a poor denoising effect. The CNN model preserved the feature information and texture structure of the original image to the maximum extent; however, the DnCNN model was inefficient at processing the edge noise.

The statistical analysis was carried out for the data at a noise level of σ =3 (*Table 5*). The Mann-Whitney U test was used, as the data did not follow a normal distribution. The p values were less than 0.05, indicating that there was a statistically significant difference.

To verify the effectiveness of our model in reducing speckle noise in real clinical ultrasound images, the private ultrasound image data set provided by Shanghai Children's Medical Center with 100 cardiac ultrasound images was tested. *Figure 6* displays the visual denoising effect of all models in this data set. Compared to other models, our model was able to effectively remove noise from real clinical ultrasound images while also providing better clarity and visual effects than other models. Moreover, as *Table 6* shows, the present denoising model outperformed the other models in terms of the objective evaluation metrics.

## Discussion

### *Summary of the experimental results*

In this study, we proposed an ultrasound image denoising autoencoder model that uses the lightweight attention mechanism to address the ineffective reduction of speckle noise and loss of detail in current deep-learning ultrasound image denoising algorithms.

Experiments on images with artificial speckle noise and real clinical ultrasound images demonstrated the superior performance of the LAD-CNN model over conventional filtering methods and other CNN models in reducing speckle noise while retaining texture structures. Based on the objective metrics and subjective visual evaluation, we observed that the conventional denoising models had the defects of over-smoothing and incomplete noise reduction. The deep-learning denoising models, such as the CNN-

 

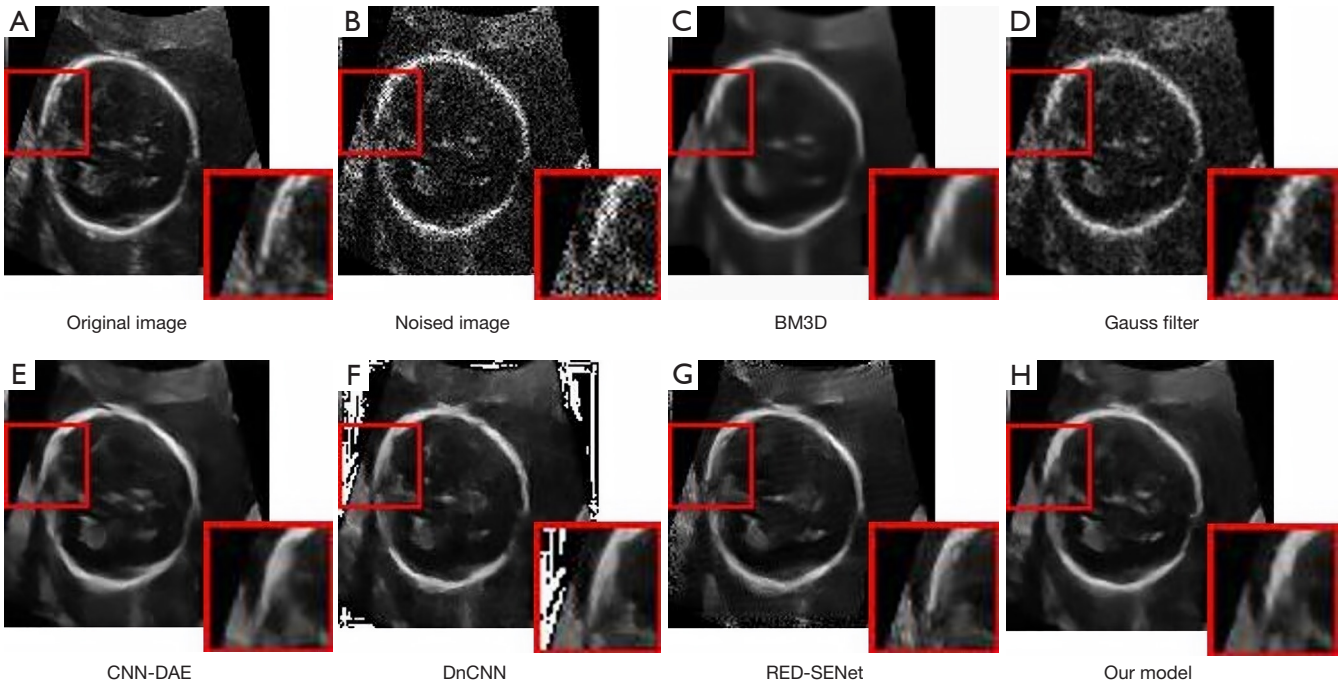


**Figure 5** Subjective visual comparison of the denoised fetal head images at σ =5.0. (A) Original image, (B) noised image, (C) BM3D, (D) Gauss filter, (E) CNN-DAE, (F) DnCNN, (G) RED-SENet, (H) our model. BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network; σ, noise level.

**Table 5** PSNR statistics for each model at the noise level of σ =3

| Model | BM3D | Gauss filter | CNN-DAE | DnCNN | RED-SENet |
|---|---|---|---|---|---|
| P value | <0.001 | <0.001 | 0.001 | <0.001 | 0.004 |

PSNR, peak signal-to-noise ratio; σ, noise level; BM3D, block-matching and 3D filtering; CNN-DAE, convolutional neural network–denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

DAE, DnCNN, and RED-SENet models, generally performed better on the objective metrics, but had the problems of blurred image texture and noisy edges on the subjective visual images. In both the brachial plexus test set and fetal head test set, the LAD-CNN model outperformed the conventional denoising models and the other deep-learning denoising models.

Further, it is worth noting that the RED-SENet model exhibited artifacts on the brachial plexus test set that were not present in (12) and other test sets in the present study. The training and test sets in (12) were from the same ultrasound data set, while the training set and test sets were from different data sets in the present study. It is speculated that the RED-SENet model may be less effective when the training and test sets are from different data sets, particularly when dealing with low-quality images. Therefore, we suggest that the same data set be used for training and testing to prevent the occurrence of artifacts on low-quality ultrasound images.

### *Potential application in medical diagnostics*

To evaluate the denoising effect of the LAD-CNN model in clinical diagnosis, eight ultrasonographers (two physicians and six technicians, all with more than five years of experience) used their expertise and experience

 

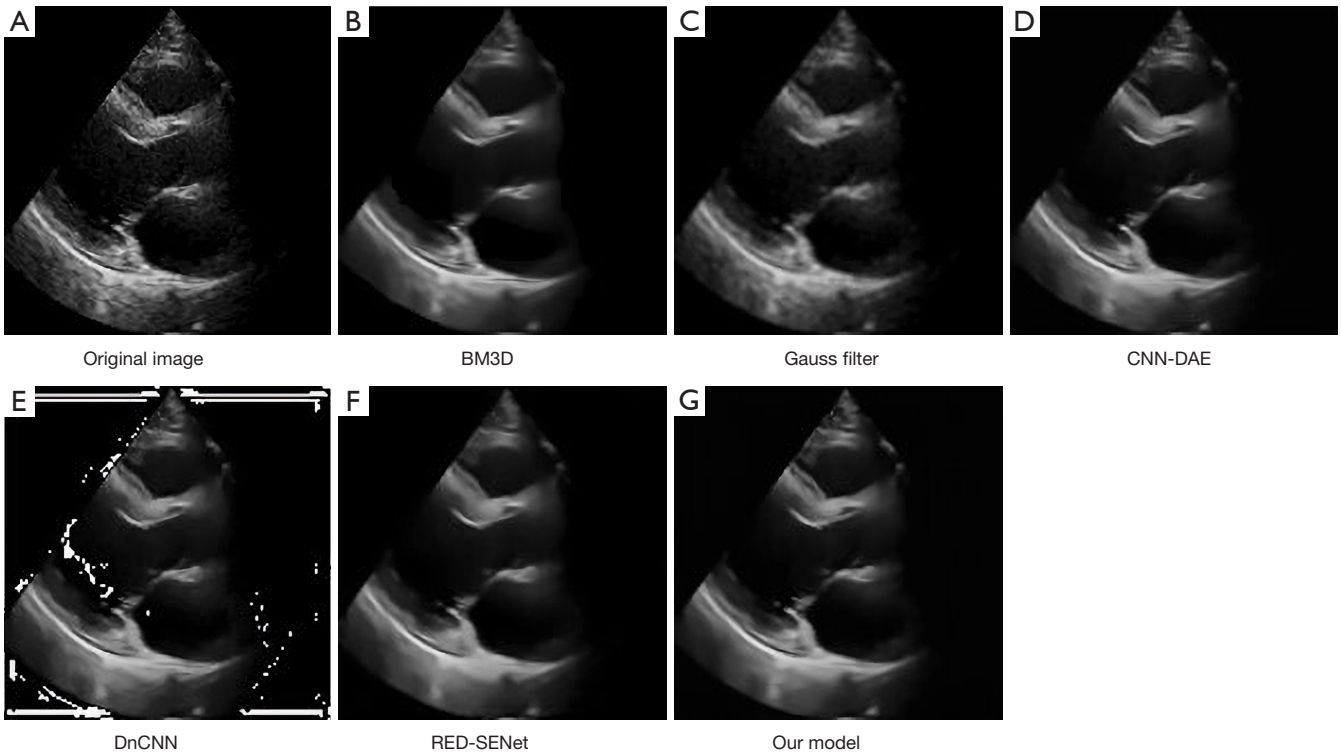


**Figure 6** Subjective visual comparison of the denoised images from different models using the private data set. (A) Original image, (B) BM3D, (C) Gauss filter, (D) CNN-DAE, (E) DnCNN, (F) RED-SENet, (G) our model. BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

**Table 6** Comparison of the metrics of the various models using the private data set

| Model | BM3D | Gauss filter | CNN-DAE | DnCNN | RED-SENet | Our model |
|---|---|---|---|---|---|---|
| ENL | 13.78 | 13.59 | 16.64 | 13.15 | 14.41 | 16.73 |

BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network; ENL, equivalent number of looks.

to evaluate 100 real ultrasound data sets. The evaluation criteria were whether the ultrasound images were clearer after denoising than before denoising, and whether it was easier to confirm the lesions or organs during diagnosis. The results of the questionnaire are shown in *Figure* 7. Six ultrasonographers were of the view that the LAD-CNN model was more effective in denoising (see *Figure* 7*A*), and seven ultrasonographers were of the view that our model could assist in ultrasound diagnosis (see *Figure* 7*B*). The results further confirmed the effectiveness of the proposed denoising model.

The time taken to process a 128 px × 128 px—sized ultrasound image was compared with the conventional model and the deep-learning models at a noise level of $\sigma$ =3.0 (see *Table* 7). Between the two conventional models, BM3D shows a better denoising effect than Gaussian filtering, but the time cost was greatly increased. In comparison, the deep-learning models had excellent denoising performances and basically realized "real-time denoising". Thus, the use of deep-learning denoising represents a promising method for image processing.

 

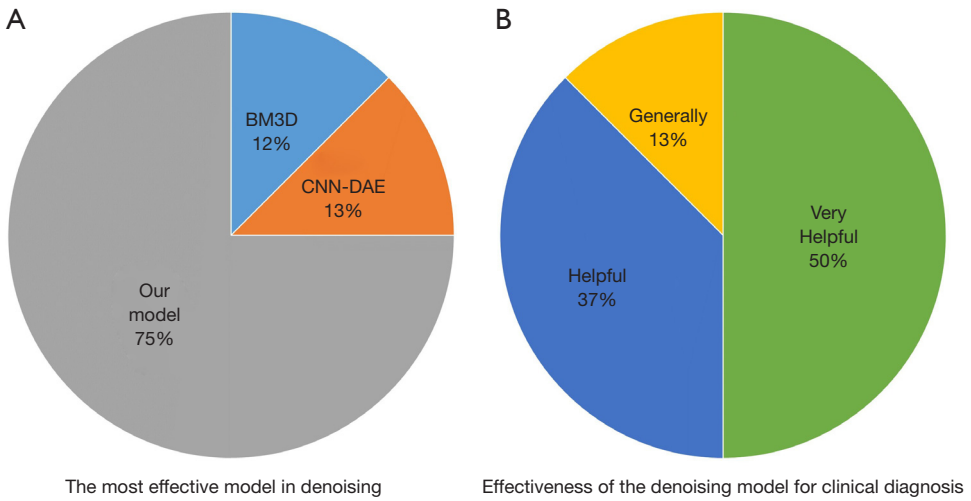


**Figure 7** Questionnaire on the denoising effect of real ultrasound images. (A) The most effective model in denoising. (B) Effectiveness of the denoising model for clinical diagnosis. BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder.

**Table 7** Comparison of denoising time

| Model | BM3D | Gauss filter | CNN-DAE | DnCNN | RED-SENet | Our model |
|---|---|---|---|---|---|---|
| Time(s) | 7.150 | 0.08 | 0.008 | 0.009 | 0.008 | 0.008 |

BM3D, block-matching and three-dimensional filtering; CNN-DAE, convolutional neural network-denoising autoencoder; DnCNN, denoising convolutional neural network; RED-SENet, residual encoder-decoder with squeeze-and-excitation network.

**Table 8** Comparison of the metrics from the ablation experiments

| Model | Noise level | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | σ =2 | | σ =3 | | σ =4 | | σ =5 | | σ =6 | | σ =7 | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Model I | 30.45 | 0.90 | 28.81 | 0.90 | 27.64 | 0.86 | 26.89 | 0.84 | 26.01 | 0.85 | 25.60 | 0.83 |
| Model II | 30.52 | 0.91 | 29.01 | 0.90 | 27.72 | 0.87 | 27.06 | 0.84 | 26.16 | 0.85 | 25.69 | 0.86 |
| Model III | 30.73 | 0.91 | 29.03 | 0.90 | 27.83 | 0.87 | 27.05 | 0.85 | 26.23 | 0.85 | 25.88 | 0.87 |
| Model IV | 30.88 | 0.92 | 29.13 | 0.91 | 28.09 | 0.88 | 27.25 | 0.86 | 26.79 | 0.86 | 26.01 | 0.85 |

σ, noise level; PSNR, peak signal-to-noise ratio; SSIM, structural similarity.

### *Ablation experiments*

To assess the effect of the two attention blocks added to the present model (i.e., the LCA block and the LLA block) on denoising performance, ablation experiments were carried out on the base model of the CNN-DAE model. As *Table 8* shows, Model I refers to the base model without adding any block, Model II incorporates a LLA block only, Model III involves a LCA block, and Model IV incorporates both lightweight attention blocks. Except for the aforementioned differences, the same settings as those for the model training were used, and 100 fetal head data sets were used as the test set.

As *Table 8* shows, the metrics of Model I were the lowest at all noise levels. The metrics increased with the addition of different blocks, and the metrics of Model IV achieved the highest values. The LCA block enabled the model to

 

**Table 9** PSNR statistics from the ablation experiments at the noise level of σ =3

| Model | Model I | Model II | Model III |
|---|---|---|---|
| P value | 0.011 | 0.020 | 0.024 |

PSNR, peak signal-to-noise ratio; σ, noise level.

adjust attention dynamically among different channels by introducing a channel attention mechanism. The LLA block expanded the receptive field by increasing the size of the convolutional kernel, enabling global information to be captured more comprehensively. Both blocks contributed to the improvement of speckle noise reduction. Mann-Whitney U tests were conducted on the data at the noise level of σ =3. The results, as shown in *Table 9*, indicate that all the P values were below 0.05, suggesting the presence of a significant difference.

## Conclusions

To address the issues of insufficient denoising performance and the loss of detail in existing deep-learning algorithms for denoising ultrasound images, we proposed the LAD-CNN model. The model incorporates a LCA block and a LLA block, which are concatenated into the encoding and decoding stages. Skip connections are employed to preserve image details and prevent overfitting. Subpixel convolutional layers are introduced in the decoding stage to enlarge the receptive field. To address the issue of artifacts, a composite loss function (comprising the linear combination of the MSE and TV functions) is employed. Comprehensive subjective visual evaluations and objective metrics were used to examine the denoising effect on both simulated speckle noise images and real ultrasound images. Comparisons against conventional denoising models and deep-learning models demonstrated that the proposed model effectively reduced speckle noise and preserved texture details in the ultrasound images.

Our future work will focus on the feasibility of practical medical applications while continuously optimizing the performance of the model. First, we plan to explore more efficient model structures to improve the performance of deep-learning models in medical image denoising. Second, we will focus on translating the results into practical medical applications and efficiently integrating deep-learning models into the medical image processing. We aim to make this preprocessing process available to physicians, enabling them to assist in diagnoses.

## Acknowledgments

*Funding:* This work was supported by the National Natural Science Foundation of China (No. 12172227).

## Footnote

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1654/rc). The authors have no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was carried out with the approval of the local institutional review board of Shanghai Children's Medical Center Affiliated to Shanghai Jiao Tong University School of Medicine, and written informed consent was obtained from the parents of the patients.



## References

1. Singh P, Mukundan R, de Ryke R. Synthetic models of ultrasound image formation for speckle noise simulation and analysis. 2017 International Conference on Signals and Systems (ICSigSys), Bali, Indonesia, 2017:278-84.
2. Rodrigues C, Peixoto ZMA, Ferreira FMF. Ultrasound image denoising using wavelet thresholding methods in association with the bilateral filter. IEEE Latin America Transactions 2019;17:1800-7.
3. Vimalraj C, Esakkirajan S, Sreevidya P. DTCWT with fuzzy based thresholding for despeckling of ultrasound images. 2017 International Conference on Intelligent

 

Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017:515-9.
4. Kasparis T, Tzannes NS, Chen Q. Detail-preserving adaptive conditional median filters. J Electron Imaging 1992;1:358-64.
5. Frost VS, Stiles JA, Shanmugan KS, Holtzman JC. A model for radar images and its application to adaptive digital filtering of multiplicative noise. IEEE Trans Pattern Anal Mach Intell 1982;4:157-66.
6. Chen K, Chen L, Xiao J, Li J, Hu Y, Wen K. Speckle reduction in digital holography with non-local means filter based on the Pearson correlation coefficient and Butterworth filter. Opt Lett 2022;47:397-400.
7. Tian C, Zheng M, Zuo W, Zhang S, Zhang Y, Lin CW. A cross Transformer for image denoising. Inf Fusion 2023;102:102043.
8. Li Q, Li S, Li R, Wu W, Dong Y, Zhao J, Qiang Y, Aftab R. Low-dose computed tomography image reconstruction via a multistage convolutional neural network with autoencoder perceptual loss network. Quant Imaging Med Surg 2022;12:1929-57.
9. Zhang Y, Hao D, Lin Y, Sun W, Zhang J, Meng J, Ma F, Guo Y, Lu H, Li G, Liu J. Structure-preserving low-dose computed tomography image denoising using a deep residual adaptive global context attention network. Quant Imaging Med Surg 2023;13:6528-45.
10. Liu JY, Yang YH. Denoising auto-encoder with recurrent skip connections and residual regression for music source separation. 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018:773-8.
11. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Trans Image Process 2017;26:3142-55.
12. Zeng X, Li Y, Gao G, Zhao X. Channel adaptive ultrasound image denoising method based on residual encoder-decoder network. Journal of Electronics & Information Technology 2022;44:2547-58.
13. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020;42:2011-23.
14. Song CH, Han HJ, Avrithis Y. All the attention you need: global-local, spatial-channel attention for image retrieval. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022:439-48.
15. Li Z, Lang C, Liang L, Zhao J, Feng S, Hou Q, Feng J. Dense attentive feature enhancement for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 2022;32:8128-41.
16. Zou W, Gao H, Chen L, Zhang Y, Jiang M, Yu Z, Tan M. Cross-view hierarchy network for stereo image super-resolution. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023:1396-405.
17. Zhong C, Gong N, Zhang Z, Jiang Y, Zhang K. LiteCCLKNet: A lightweight criss-cross large-kernel convolutional neural network for hyperspectral image classification. IET Comput Vis 2023;17:763-76.
18. Li H, Nan Y, Del Ser J, Yang G. Large-kernel attention for 3D medical image segmentation. Cogn Comput 2023. doi: 10.1007/s12559-023-10126-7.
19. Gondara L. Medical image denoising using convolutional denoising autoencoders. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016:241-6.
20. Argenti F, Torricelli G. Speckle suppression in ultrasonic images based on undecimated wavelets. EURASIP J Appl Sig P 2003;5:470-8.
21. Loupas T, McDickenW N, Allan P L. An adaptive weighted median filter for speckle suppression in medical ultrasonic images. IEEE Transactions on Circuits and Systems 1989;36:129-35.
22. Lan Y, Zhang X. Real-time ultrasound image despeckling using mixed-attention mechanism based residual UNet. IEEE Access 2020;8:195327-40.
23. Yu H, Ding M, Zhang X, Wu J. PCANet based nonlocal means method for speckle noise removal in ultrasound images. PLoS One 2018;13:e0205390.
24. Perera MV, Bandara WG, Valanarasu JM, Patel VM. Transformer-Based SAR Image Despeckling. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium 2022;751-4.
25. Wang Z, Chen J, Hoi SCH. Deep Learning for Image Super-Resolution: A Survey. IEEE Trans Pattern Anal Mach Intell 2021;43:3365-87.
26. Guo MH, Lu CZ, Liu, ZN, Cheng MM, Hu SM. Visual attention network. Comp Vis Media 2023;9:733-52.
27. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, 2001;2:416-23.
28. Wang Y, Geng J, Zhou C, Zhang Y. Segmentation

 

of ultrasound brachial plexus based on U-Net. 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021:482-5.
29. van den Heuvel TLA, de Bruijn D, de Korte CL, Ginneken BV. Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One 2018;13:e0200412.
30. Parhad SV, Aher SA, Warhade KK. A comparative analysis of speckle noise removal in SAR images. 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021:1-4.
31. Ko J, Lee, S. SAR Image Despeckling Using Continuous Attention Module. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2022;15 3-19.
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600-12.
33. Dabov K, Foi A, Katkovnik V, Egiazarian K. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans Image Process 2007;16:2080-95.
34. Wang P, Zhang H, Patel VM. SAR image despeckling using a convolutional neural network. IEEE Signal Proc Let 2017;24:1763-7.

**Cite this article as:** Shi L, Di W, Liu J. Ultrasound image denoising autoencoder model based on lightweight attention mechanism. Quant Imaging Med Surg 2024;14(5):3557-3571. doi: 10.21037/qims-23-1654

 