IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 October 2024, accepted 6 November 2024, date of publication 13 November 2024, date of current version 22 November 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3496907*

RESEARCH ARTICLE

# Speckle Noise Reduction for Medical Ultrasound Images Using Hybrid CNN-Transformer Network

**ANPARASY SIVAANPU[1], (Graduate Student Member, IEEE), KUMARADEVAN PUNITHAKUMAR[1], (Senior Member, IEEE), RUI ZHENG[2], (Member, IEEE), MICHELLE NOGA[1], DEAN TA[1,3,5], (Senior Member, IEEE), EDMOND H. M. LOU[4], (Senior Member, IEEE), AND LAWRENCE H. LE[1,5]**

[1]Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, AB T6G 1H9, Canada
[2]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
[3]Department of Biomedical Engineering, School of Information Science and Technology, Fudan University, Shanghai 200437, China
[4]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
[5]State Key Laboratory of Integrated Chips and Systems, Fudan University, Shanghai 200437, China

Corresponding author: Lawrence H. Le (lawrence.le@ualberta.ca)

This work was supported in part by the Alberta Innovates—Accelerating Innovations into CarE (AICE) Concepts; in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance—Alberta Innovates Programs; in part by the Discovery Grant from NSERC; in part by the Natural Science Foundation of China (NSFC) under Grant 12074258; and in part by the Senior Visiting Scholarship from the State Key Laboratory of Integrated Chips and Systems, Fudan University, Shanghai, China. The work of Anparasy Sivaanpu was supported by Alberta Innovates Graduate Fellowship.

**ABSTRACT** Ultrasound images are often affected by limited resolution, artifacts, and inherent speckle noise. To address these challenges, researchers have explored denoising approaches. Recently, deep learning methods have demonstrated distinct advantages in ultrasound image denoising. However, further improvements are needed to preserve structural details, such as boundaries, edges, and margins. This paper proposes a hybrid CNN-transformer network called HCTSpeckle, an encoder-decoder network with a fusion block designed to enhance ultrasound images. The fusion block combines swin transformers to capture global modeling relationships, and convolutional neural networks to extract local modeling details. It is integrated into the encoder-decoder structure, allowing the model to focus on both local and global texture structural information. An improved swin block is also introduced into the network to improve robustness by extracting more significant features. HCTSpeckle was evaluated both quantitatively and qualitatively with clinical objectives using two public and two private datasets. Both results showed that HCTSpeckle significantly enhanced the ultrasound image quality and outperformed state-of-the-art methods in noise reduction and structure preservation across all four datasets. Compared to existing denoising methods, HCTSpeckle achieved notably faster performance in terms of complexity comparison, such as parameter counts, gigaFLOPs, and inference time. Moreover, this study assessed the effectiveness of HCTSpeckle for alveolar bone segmentation using dental images, demonstrating that HCTSpeckle significantly improved segmentation performance. Furthermore, an experienced radiologist blindly rated the 250 dental US images on a scale of 1 to 5, with 5 being the highest image quality, showing that HCTSpeckle consistently produced higher-quality images.

**INDEX TERMS** Convolutional neural network, deep learning, hybrid network, image denoising, intraoral ultrasound, speckle noise, supervised learning, ultrasound imaging.

## I. INTRODUCTION

Medical ultrasound (US) imaging has been widely used in diagnostic applications because of its low cost, noninvasive nature, safety, portability, and real-time capability compared

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

to other modalities. An inherent characteristic of US images is speckle noise, which is caused by backscattered signals, and exhibits multiplicative noise features with a signal dependent distribution [1]. The US speckle noise model is governed by the distribution of Gamma [2] or Fisher-Tippett [3] and represented by

$$J(x, y) = I(x, y) + I^{\gamma}(x, y)\eta(x, y) \tag{1}$$

where $J(x, y)$ is the noisy image, $I(x, y)$ denotes a noise-free image; $\eta(x, y)$ is the Gaussian noise with zero mean and variance $\sigma^2$; $x$ and $y$ are the spatial coordinates of the variables. The variable $\gamma$ is constant depending on US devices and imaging processing, and is set to 1 to define the multiplicative speckle noise model [4], [5]. Speckle noise is inevitable in US images and degrades visual quality. Furthermore, the presence of speckle noise makes it challenging for clinicians to diagnose lesions accurately because it hinders the extraction, analysis, and recognition of lesion characteristics [6]. Therefore, reducing noise while preserving the structural details in sonograms is crucial for improving image quality for more accurate diagnostic information.

Over the past few years, several approaches have been proposed to mitigate or reduce speckle noise in US images using traditional techniques such as spatial domain [7], transform domain [8], and hybrid domain filtering [9]. Applying a spatial filter directly to the image yields a lower computational complexity [10] and reduces noise but leads to increased image blurring. Transform-domain techniques convert an image from one domain to another, leveraging distinct image properties in the transform domain for denoising [11]. Hybrid filtering, which integrates multiple techniques, offers effective noise reduction and edge preservation [12]. However, their computational complexity remains relatively high and limits their application owing to the real-time processing demands of US image applications. Although these methods aim to decrease speckle noise and improve image quality, determining the optimal trade-off between smoothing and preserving image details remains a challenge. Determining the ideal balance between denoising performance and algorithm complexity is even more problematic, particularly when handling real-time US 3D data. Consequently, researchers are actively exploring methods with lower algorithmic complexity to effectively remove speckle noise in US images as post-processing methods. The post-processing approach for denoising US images is advantageous because of its independence from projection data, high portability, and user-friendliness. However, conventional techniques do not fully address the issues related to excessive smoothing and the introduction of new noise components into the image resulting from the processing.

Unlike traditional methods, artificial intelligence (AI), particularly deep learning (DL), has been used in numerous applications such as classification [25], segmentation [26], detection [27], super-resolution [28], and denoising [29]. DL offers a promising and valuable avenue for real-time and effective despeckling of US images [30] by learning and mapping the intrinsic features of training samples. The authors introduced a range of medical image-denoising techniques rooted in DL, beginning with the development of denoising models and refining loss functions. Generative adversarial networks (GANs) [31], [32], [33], [34], [35], [36], [37] and convolutional neural networks (CNNs) [29], [38], [39], [40], [41], [42] were used to produce high-quality noise-free US images in an end-to-end manner. Approaches such as fast and flexible CNN [41] and flexible denoising CNN [43] require noise estimations and extensive training data. Furthermore, a notable framework for image denoising is the convolutional encoding and decoding approach, which has demonstrated impressive performance. For instance, an encoder-decoder with a residual learning technique [44] has shown promising results in image restoration. Similarly, a successive encoder-decoder network [45] utilized symmetric skip connections for image denoising and super-resolution. In [46], a deep CNN was combined with an autoencoder and skip connections, to introduce a shallow residual encoder-decoder network to denoise the images. A residual encoder-decoder, known as the Wasserstein generative adversarial network (RED-WGAN) was introduced in [47] to denoise 3D MRI images. However, the performance of these CNN models is highly dependent on accurate noise level estimation, which leads to poor performance when the noise level is unknown. Furthermore, training CNN models from scratch is time-consuming, computationally expensive, and requires extensive data for training purposes.

Although well-known dimension reduction CNN techniques are highly effective in extracting valuable features from an image, attention mechanisms have recently shown great success in many computer vision tasks, including image restoration and enhancement. The attention module generally includes both the channel and spatial attention mechanisms. These attention maps were applied to the feature map to refine the features adaptively [23]. Integrating spatial attention into CNNs facilitates the capture of spatial correlations and effectively modeling spatial dependencies between features [48]. The residual encoder-decoder with a squeeze and excitation network (REDSENet) [24] introduced a denoising model that operates in channels and effectively mitigates speckle noise from US images. The REDSENet model is built upon a residual encoder-decoder architecture, and by incorporating an attention block within the decoder section of the given model, it can acquire and leverage global information. In [49], the authors proposed a separation and refusion strategy for the attention mechanism, which was fed into the encoder phase of the standard UNet architecture [50] to perform real-time despeckling. Another implementation of the attention module with the residual encoder-decoder is RED-MAM [23]. The RED-MAM network incorporated multiattention mechanisms in its decoding phase. The output features were combined with the corresponding encoded

**TABLE 1. Benefits & limitations of US denoising methods.**

| Methods | Benefits | Limitations |
|---|---|---|
| Spatial domain based [7], [13]–[15] | 1. Less complexity | 1. Image blurring<br>2. Loss of image details<br>3. Degradation of resolution |
| Transform domain based [16], [17] | 1. Good edge preservation capability | 1. Susceptibility of creating artifacts<br>2. High complexity |
| Hybrid based [9], [18] | 1. Good denoising capability<br>2. Good edge preservation capability | 1. Substantial processing time<br>2. High complexity |
| CNN-based [4], [19]–[22] | 1. Flexible network structure<br>2. Suitability for real-time use | 1. Smaller receptive field<br>2. Insufficient distinction in channel features |
| Attention-based [23], [24] | 1. Good denoising capability<br>2. Focus of critical details<br>3. High robustness | 1. High computational requirements |

features, and fed to the next level. Multiattention enables the model to highlight essential content features within key channels while minimizing the influence of less relevant features. However, the addition of attention modules to the encoding and decoding phases led to unsharp edges, thereby diminishing the overall effectiveness of the denoising model. In summary, Table 1 presents a thorough comparison of the advantages and disadvantages of the denoising techniques, offering their strengths and weaknesses.

Following CNN and attention mechanisms, transformers [51] have emerged in natural language processing and have recently gained an attraction in computer vision. These advances have now been extended to medical imaging, where transformers have proven successful in various medical applications, such as image reconstruction, segmentation, detection, and diagnosis [52]. In [53], researchers proposed a TED-net, an encoder-decoder dilation network combined with a token-to-token (T2T) vision transformer for low-dose CT denoising. The TED-net employs a U-shaped model and dilation during the T2T phase to expand the receptive field. In [54], transformers were harnessed in medical image denoising with an edge enhancement called Eformer, which utilizes transformer blocks to construct an encoder-decoder framework designed explicitly for CT imaging. Despite recent advancements in transformer-based denoising techniques, there are still obstacles to the effective denoising of US images, particularly simultaneous noise removal while preserving the structural details.

To address the problems of using either CNN or transformer techniques for US image denoising, we introduced the CNN-transformer hybrid network, an encoder-decoder denoising network with a swin transformer and a residual CNN (res) block, which is called HCTSpeckle. The strengths of the CNN and transformer were combined into a fusion block and used as the backbone of HCTSpeckle. To the best of our knowledge, this is the first hybrid CNN-transformer-based denoising network specifically developed for US imaging. We trained the proposed HCTSpeckle with multiple noise levels to perform better on US images with intricate noises. The significant contributions of this study are as follows:

- The HCTSpeckle model for US image denoising is designed to extract richer features from US images by preserving local and global modeling details with less computational cost.
- The proposed fusion block combines a res block and swin transformer block to avoid feature loss and preserve long-range dependencies locally and globally.
- An improved swin block is introduced into the network to improve the robustness of the proposed denoiser by extracting more significant features, thereby enhancing the noise-removal effectiveness.
- HCTSpeckle is validated through extensive experiments on real and synthetic US datasets, such as two private and two public datasets with multiple noise levels. The results of HCTSpeckle show that it performs faster and better than other state-of-the-art techniques, both qualitatively and quantitatively.
- In addition, the performance of alveolar bone segmentation with and without denoising by HCTSpeckle is tested using dental US data. The segmentation metrics prove that HCTSpeckle improves segmentation performance. Furthermore, an experienced radiologist blindly compared and rated the denoised images for image quality, contrast, resolution, and noise. Moreover, statistically significant differences are determined between the HCTSpeckle and the most recent best-performing methods. The obtained results are favourable for the proposed HCTSpeckle technique.

The remainder of this article is organized as follows: A review of related work is presented in Section II; a detailed description of HCTSpeckle's overall structure is presented in Section III; the experimental setup is described in Section IV; the experimental results, computational complexity comparison, quantitative results of segmentation performance, blinded qualitative validation, and statistical analysis of the results are discussed in Section V. Finally, Section VI concludes our study.

## II. RELATED WORK

### A. CNN-BASED METHODS

Recent advancements in DL have led to the development of innovative image denoising techniques. DL-based approaches, particularly CNNs, utilize end-to-end architectures with powerful learning capabilities and have been extensively used in US image denoising. Encoder-decoder networks with skip connections in both the convolutional and transposed convolutional layers are well known for their effectiveness in image denoising tasks. This design allows for the extraction of finer features from the bottom layer, thereby enhancing denoising performance. References [23], [24], and [49] employed CNN encoder-decoder design to develop effective denoising networks. Although CNN models extract features using stacked, uniformly sized kernels, they often suffer from a loss of texture detail and cause over-smoothing in US images. Additionally, CNN approaches have inherent limitations in US denoising due to biases in convolutional operations. Enhancing these models typically requires greater computational resources.

### B. TRANSFORMERS

The transformer, an alternative to CNNs, employs a self-attention mechanism to capture global interactions within contexts and has demonstrated promising performance in various vision tasks. Vision transformers have been used for image restoration [55], [56], but they typically face significant computational challenges due to the quadratic complexity of the self-attention mechanisms. Recently, swin transformers [57] have shown significant promise in image restoration tasks, with some extensions being made for real-world image denoising [58]. However, these approaches incur high computational costs due to the extensive use of transformer blocks.

Recent DL-based approaches for US image denoising predominantly rely on either CNNs or transformer techniques. Each method has its strengths, i.e., CNNs are efficient at local feature extraction, while transformers excel at capturing global interactions. However, both approaches have limitations when used in isolation. To overcome these challenges and achieve superior denoising performance for US images with lower computational costs, we introduced HCTSpeckle, a hybrid network that combines the strengths of CNNs and swin transformers. By integrating these two architectures, HCTSpeckle effectively balances the benefits of both local and global feature extraction, leading to improved denoising results for US images.

## III. THE PROPOSED METHOD

In this section, detailed information on the proposed HCTSpeckle method is provided. An encoder-decoder based denoising network with a fusion of a swin transformer and a res block to focus on local and global features without dependency loss was designed for US image denoising to improve US image quality. The design of the proposed method was justified using a comprehensive ablation study.

### A. NOISE MAPPING DL MODEL

DL offers an effective solution for denoising US images by learning from data samples to model noise instead of making changes in hardware resources such as transducer properties. The noise mapping between the noisy image, $J \in R^{M\times N}$ and the noise-free image, $I \in R^{M\times N}$ can be expressed by

$$J = \Omega(I) \tag{2}$$

where $\Omega$ is the mapping function of the noise distribution. Using (2), the denoising approach determines the best approximation of $\Omega^{-1}$. An overview of the US image-denoising process is given by

$$\underset{\rho}{\operatorname{argmin}} \|\hat{J} - I\|_2^2 \tag{3}$$

where $\hat{J} = \rho(J)$ is the estimation of $J$, $\rho$ denotes the optimal approximation of $\Omega^{-1}$. Noisy and corresponding noise-free image pairs are required to train the DL-based model in a supervised manner. In US imaging, it is challenging to obtain a noise-free US image as the noise in US imaging is dependent on many properties such as tissue inhomogeneities and the device used for acquisition. Therefore, the original US data are considered as reference noise-free images that underwent physical correction by adjusting acquisition parameters, resulting in acquired images with minimal real noise. The various amounts of speckle noise is added to the acquired data using noise simulation technique and used as noisy images. These sets of data served as reference noise-free and noisy images for training purposes. Publicly available US datasets inherently contain noise, which is typically retained and used to generate more noisy data, which is used as a pair of reference noise-free and noisy data for training purposes. The training procedure of the DL network is represented by

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \frac{1}{N} \sum_i ||f(I_i, n_i; \Theta) - I_i|| \tag{4}$$

where $I_i$ is the $i^{th}$ reference noise-free image in the training stage, $n_i$ denotes the noise, $f(I; \Theta)$ represents the denoising network, $\Theta$ represents the network weights, and $N$ denotes the total number of training images.

### B. THE PROPOSED HCTSPECKLE ARCHITECTURE

The overall structure of the proposed HCTSpeckle network is shown in Fig. 1. The proposed methodology comprises four levels of encoding blocks and an additional four decoding blocks, as shown in Fig. 2. The initial feature ($F$) is extracted from the input US image by using the first convolutional layer. Subsequently, $F$ is used as the input for the encoding layer through the fusion block as the Swin-Res (SR) block. This fusion block is the backbone of the proposed HCTSpeckle to achieve better denoising performance. The architecture of an SR block is shown in Fig. 1. Between each encoding and decoding layer, an SR block is introduced to enhance feature extraction on a multi-resolution scale.

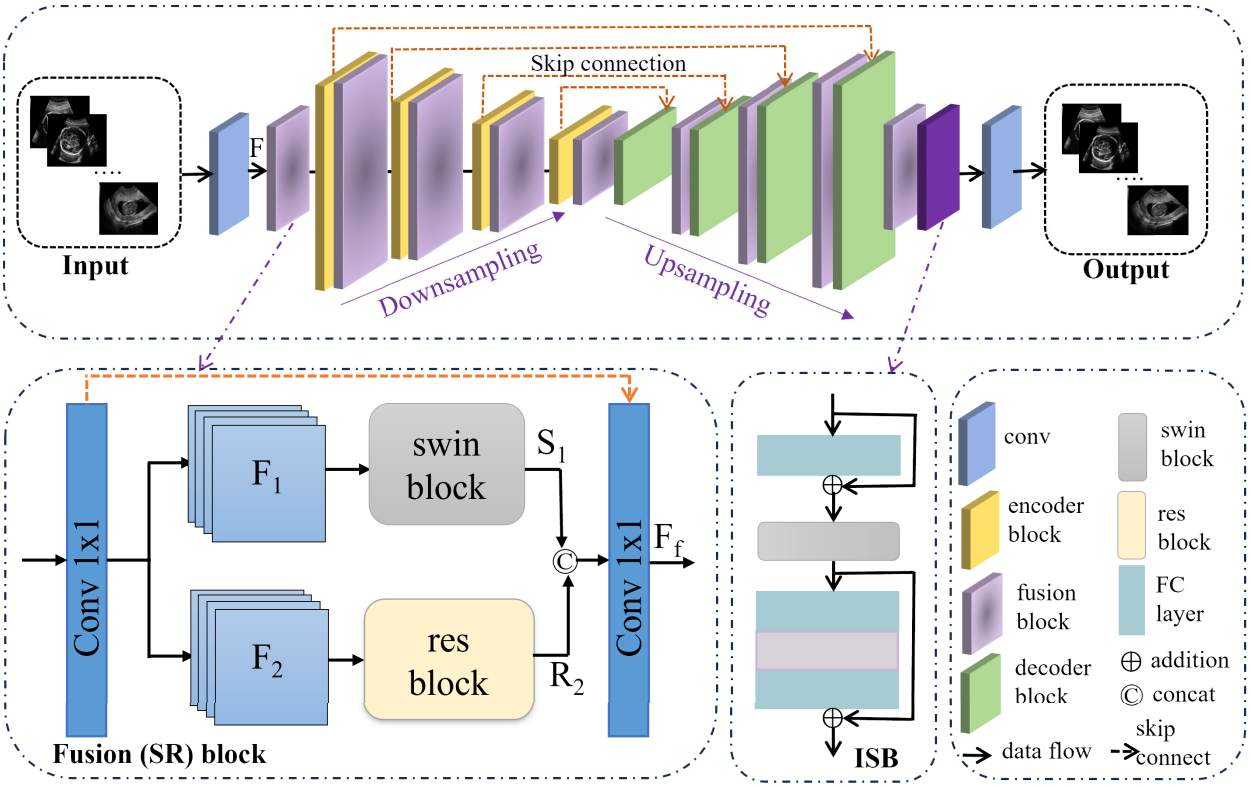



**FIGURE 1.** The structure of the proposed HCTSpeckle denoising network. The input 2D US image is passed through the proposed architecture and the denoised image is obtained as the result of the network. Encoder, decoder block, fusion block (SR) and improved swin block (ISB) are depicted by yellow, green, light purple and dark purple boxes, respectively. The skip connections are indicated by orange lines. FC layer and res block represent fully connected layer and residual CNN block, respectively. The structure of fusion block and improved swin block are shown below in the block diagram.

In addition, skip connections are incorporated into the HCTSpeckle model to accelerate the training process while preserving the finer details. Furthermore, the improved swin block is integrated with the encoder-decoder network to improve the robustness by extracting more significant features, thereby enhancing the noise-removal effectiveness. Finally, a single convolutional layer is used to produce the final output, i.e., a denoised image with enhanced visual quality and clear structural details.

### 1) ENCODER-DECODER STRUCTURE

The proposed HCTSpeckle model consists of four encoding and decoding blocks that are interconnected through SR blocks. Each encoder block uses a $2 \times 2$ strided convolution with a stride of two, whereas the decoder block uses a transposed convolution. The rectified linear unit (ReLU) activation function is applied to each encoding and decoding layer. The number of channels in each block varies from $64 \times 64$ in the first scale to $128 \times 128$ in the second, $256 \times 256$ in the third, and $512 \times 512$ in the fourth, as shown in Fig. 2.

### 2) FUSION (SR) BLOCK

The internal structure of a fusion (SR) block is illustrated in Fig. 1. The SR block combines a swin transformer (swin) block and res block.

The swin transformer is constructed by substituting the standard multi-head self-attention (MHSA) mechanism in a transformer block with shifted windows while keeping the others unchanged [57]. The swin block enhances computational efficiency and improves the model's ability to capture long-range dependencies compared to traditional self-attention mechanisms. Within each window, the block utilizes MHSA modules along with multi-layer perceptron (MLP) for further non-linear transformations. Each MHSA mechanism and MLP is preceded by a layer normalization (LN) layer, and a residual connection is added after each module. LN stabilizes the training process and the residual connection supports a better gradient flow. These combined features allow the swin block to effectively balance local detail extraction and global context understanding, making it suitable for image denoising.

The res block combines both linear and non-linear elements within the CNN architecture to extract more comprehensive features. The convolutional layer is responsible for extracting linear features, whereas the ReLU functions as piecewise activation to convert these linear features into non-linear ones. Considering the long-term dependency issue and robustness of the obtained structural information, the linear and non-linear features from the first and third layers are fused through a residual learning operation, serving as the input for the concatenation operation with a swin block.

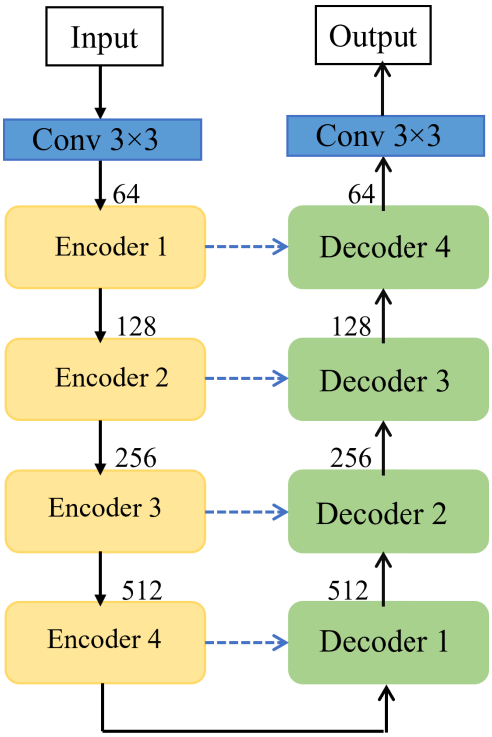



**FIGURE 2.** The structure of the proposed encoder-decoder network. Encoder and decoder are depicted by yellow and green blocks, respectively.

Initially, the input feature ($F$) is passed through $1 \times 1$ convolution and evenly split into two groups of features ($F_1$ and $F_2$), which are then fed into the swin and res blocks in parallel. This can be stated by

$$F_1, F_2 = Split(H_{FE_{1\times1}}(F)) \tag{5}$$

where $H_{FE_{1\times1}}(.)$ represents the $1 \times 1$ convolution. $S_1$ and $R_2$ represent the results for the swin and the res blocks, respectively and are expressed by

$$S_1 = swinblock(F_1), \tag{6}$$
$$R_2 = resblock(F_2). \tag{7}$$

Subsequently, both output features ($S_1$ and $R_2$) are fused using concatenation, followed by convolution. The residual connection uses $1 \times 1$ convolution to produce the residual component of the input. The outcome of the SR block is given by

$$F_f = H_{FE_{1\times1}}(S_1 \odot R_2) + F \tag{8}$$

where $H_{FE_{1\times1}}(.)$ represents the $1 \times 1$ convolution and $\odot$ denotes the concatenation. The proposed SR block offers several advantages for US denoising. The SR block effectively combines localized modeling of the res block with non-localized modeling of the swin block. HCTSpeckle enhances local and non-local modeling capabilities by integrating a multiscale encoder-decoder architecture. The features are split into two groups, and concatenating the two groups further reduces the computational complexity and the parameter count of the proposed SR block.

#### 3) IMPROVED SWIN BLOCK

An Improved Swin Block (ISB) is used after the $9^{th}$ fusion block to remove the interference from previous interactions. The ISB comprises a sequence of fully connected (FC) layer, swin block, FC layer with ReLU activation, and a single FC layer. To enhance information acquisition, a residual learning operation is applied between the ISB input and output of the first FC layer, as well as between the swin block and FC outputs. By incorporating residual learning and multiple fully connected layers, the ISB improves the model's ability to remove noise without sacrificing important image details. The illustrations can be represented by

$$F_{isb} = FC(R(FC(S(FC(F) + F)))) + S(FC(F) + F) \tag{9}$$

where $F$ is the input feature of the proposed improved swin block, $FC$ is the FC layer, $R$ is the rectified linear unit activation function, $S$ is the swin transformer function and $F_{isb}$ is the output of the proposed $ISB$.

### C. LOSS FUNCTION

$L_1$ pixel loss is used as the loss function in the proposed HCT-Speckle model. $L_1$ loss determines the absolute pixel-wise differences between the output denoised US image ($J_{out}$) and reference noise-free US image ($I_{ref}$).

$$Loss = \frac{1}{N}\sum_{i=1}^{N} \|I_{ref(i)} - J_{out(i)}\|_1 \tag{10}$$

where $N$ denotes the number of samples.

### D. TRAINING PROCEDURE

In hyperparameter tuning, the best values for batch size, filter size, block quantity, layer count within each block, and learning rate adjustments are obtained by training for a set number of iterations. These hyperparameters are chosen based on experimental results to optimize the performance of the proposed technique. A validation loss metric is used to determine the most effective model. In each training iteration, the loss is calculated using (10). The training procedure for the HCTSpeckle model is shown in Algorithm 1.

## IV. EXPERIMATAL SETUP

### A. IMPLEMENTATION DETAILS

The proposed HCTSpeckle network was implemented using Python 3.9, and the PyTorch platform along with CUDA 11.8 was executed on a machine running Windows 11 $\times$ 64 OS with an Intel(R) Core™ i7-13700F processor and 16GB RAM. A $256 \times 256$ pixel size was used for the input block during network training. The Adam optimizer was used with an initial learning rate of $1 \times 10^{-4}$, which gradually decreased approximately to $1 \times 10^{-5}$ by the end of the training, and the training process was run for 200 iterations with a batch size of 16. The summary of the proposed model's parameter setting

**Algorithm 1** Training Procedure for HCTSpeckle

**Input:** $TrD = \{(J_1, I_1), (J_2, I_2), \ldots, (J_N, I_N)\}$, where each pair $(J_x, I_x)$ consists of a noisy US image $J_x$ and a noise-free US image $I_x$, $TrD$ denotes the training dataset.
**Output:** HCTSpeckle denoising model $\phi_{\text{denoise}}$.
**initialize:** the number of epochs $n$; batch size $b$; and learning rate $lr$.
**Detailed Process:**

**while** $i < n$ **do**
  Randomly visits the US training dataset.
  $(J_x, I_x) \leftarrow f_{data}(TrD)$ // Pairs of US images were randomly selected from $TrD$.
  **for** $l \in \{1, 2, \ldots, \frac{N}{b}\}$ **do**
    $J_l = \{J_x^l\}_{x=(l-1)b+1}^{lb}$, $I_l = \{I_x^l\}_{x=(l-1)b+1}^{lb}$ // $l^{th}$ batch of noisy US images $J_l$, the $l^{th}$ batch of noise-free US images $I_l$, the $x^{th}$ of the $l^{th}$ batch noisy US image $J_x^l$, the $x^{th}$ of the $l^{th}$ batch noise-free US image $I_x^l$.
    $D_l \leftarrow \phi_{denoise}(J_l)$ // Input $J_l$ into the denoising model $\phi_{denoise}$ to obtain the denoised image.
    $Loss \leftarrow \frac{1}{b}\sum_{x=1}^{b} \|I_x^l - D_x^l\|$ // Compute the training loss of the $l$-th batch after denoising model $\theta_d$.
    $\frac{\partial Loss}{\partial \theta_d} \leftarrow \nabla_{\theta_d} Loss$ // Calculate the gradient of $\theta_d$.
    $\theta_d \leftarrow Adam(\theta_d, lr)$ // Update the parameters $\theta_d$.
  **end for**
**end while**
Save the model after training.

is tabulated in Table 2. The testing was conducted using the parameters saved in the $200^{th}$ epoch. The rating results were statistically analyzed using IBM SPSS-27 Statistics software.

### B. DATASET DETAILS

We used two public and two private data sets to evaluate the efficacy of HCTSpeckle by using real and synthetic US images. The data details are listed in Table 3, and example images are shown in Fig. 3. The dataset details are described as follows:

(1) **Fetal head dataset [59] (HC18):** This publicly available dataset contains US images of the fetal head. During pregnancy, US imaging is used to measure fetal biometrics, including head circumference (HC), which helps estimate gestational age and monitor fetal growth. The HC measurement is obtained from a cross-sectional view of the fetal head.
(2) **Breast US images [60] (BUSI):** This publicly available dataset comprises breast US images aimed at detecting breast cancer using US scans. It includes data from 600 female patients, encompassing various aspects of breast pathology, to offer a comprehensive analysis of breast health.
(3) **Dental US images [26]:** The images were acquired in an animal study at our institution. This dataset, derived from two porcine samples, includes dental structures such as alveolar bone, enamel, and gingiva for dentoperiodontal analysis. The dental US data acquisition was performed using a 20-MHz 1D-linear US probe transducer. The imaging frame rate and depth were 18 fps and 12 mm, respectively.
(4) **Heart phantom:** US images of the heart phantom were acquired at our institution using an X5-1 matrix array transducer system. The imaging parameters were 38 fps and 17-cm imaging depth.

The original US dataset was collected by gathering image data with minimal noise through a physical correction process. To assess the robustness and stability of the proposed HCTSpeckle, speckle noise was simulated on these original US images according to the speckle noise distribution [4], [49] using (1). A constant variable $\gamma = 0.5$ was used as specified in [5] and [49]. This allows for controlled noise levels in the US dataset, enabling multiple experiments with pairs of reference noise-free and noisy data. To train the proposed neural network, the dataset was corrupted with various levels of speckle noise [0.1, 0.25, 0.5, 0.75]. Additionally, the training and test sets for the four datasets listed in Table 3 were randomly split at a 7:3 ratio from images with simulated speckle noise. By adding the noise to US datasets, the proposed HCTSpeckle was tested for its robustness under various conditions and anatomical structures.

**TABLE 2. Summary of the proposed model parameter settings.**

| Parameter | Values |
|---|---|
| Input Size | $256 \times 256$ |
| Activation | ReLU |
| Batch Size | 16 |
| Learning Rate (initial) | 0.0001 |
| Optimizer | Adam |
| Loss Function | $L_1$ loss |
| Number of Epochs | 200 |

### C. EVALUATION CRITERIA

Two sets of evaluation metrics were used for both the quantitative and qualitative purposes. Quantitative metrics

**TABLE 3. Details of datasets.**

| Datasets | Availability | Number of images | Number of training sets | Number of testing sets |
|---|---|---|---|---|
| Fetal-head US | Public | 1334 | 934 | 400 |
| Breast US | Public | 744 | 516 | 228 |
| Dental US | Private | 1500 | 1050 | 450 |
| Heart phantom | Private | 1000 | 700 | 300 |

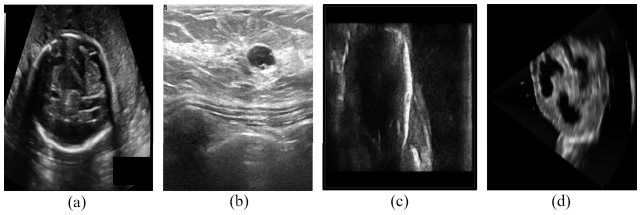



**FIGURE 3. Example image from each dataset: (a) Fetal head, (b) Breast US, (c) Dental US, and (d) Heart phantom.**

were further classified into reference metrics, namely the structural similarity index (SSIM) [61], peak signal-to-noise ratio (PSNR) [62], mean squared error (MSE), and non-reference metrics, namely speckle index (SI) [63], natural image quality evaluator (NIQE), inherent signal-to-noise ratio (ISNR) [64], entropy, contrast-to-noise ratio (CNR) [65], and signal-to-noise ratio (SNR) [66]. We assumed the reference noise-free image, denoised image, and image size be $I_{ref}$, $J_{out}$, and $(M \times N)$, respectively.

### 1) REFERENCE METRICS

MSE can be defined as

$$MSE = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} [J_{out}(x, y) - I_{ref}(x, y)]^2. \quad (11)$$

PSNR can be measured by

$$PSNR = 10 \log_{10} \left( \frac{MAX^2 I_{ref}}{MSE} \right). \quad (12)$$

SSIM can be expressed as

$$SSIM = \frac{(2\mu_{J_{out}} \mu_{I_{ref}} + X_1)(2\sigma_{J_{out} I_{ref}} + X_2)}{(\mu_{J_{out}}^2 + \mu_{I_{ref}}^2 + X_1)(\sigma_{J_{out}}^2 + \sigma_{I_{ref}}^2 + X_2)} \quad (13)$$

where $\mu_{J_{out}}$ and $\mu_{I_{ref}}$ are the means of the denoised and reference noise-free images respectively. $\sigma_{J_{out}}^2$ and $\sigma_{I_{ref}}^2$ are the variances of the recovered and the reference noise-free US images, respectively. $\sigma_{J_{out} I_{ref}}$ is the covariance between $J_{out}$ and $I_{ref}$. $X_1$ and $X_2$ are the variables used to stabilize the division. $X_1$ and $X_2$ are derived from

$$X_1 = (C_1 P)^2, X_2 = (C_2 P)^2 \quad (14)$$

where variables $C_1$ and $C_2$ are constants, and $P$ is the dynamic range of the pixel values.

### 2) NON-REFERENCE METRICS

$E(J)$ and $E(J_{out})$ denote the means of the noisy input and denoised images, respectively. Similarly, $V(J)$ and $V(J_{out})$ are the variances of noisy and denoised images, respectively. The entropy of a denoised image is expressed by

$$Entropy(J_{out}) = -sum(x. * \log(x)) \quad (15)$$

where $x$ denotes the histogram count of the pixel. The speckle index (SI) can be computed by

$$SI = \frac{\sqrt{V(J_{out})}}{E(J_{out})} \quad (16)$$

and the inherent signal-to-noise ratio (ISNR) by

$$ISNR = \frac{[E(J_{out})]^2}{V(J_{out})}. \quad (17)$$

The Contrast-to-noise ratio (CNR) can be expressed by

$$CNR = 10 \log_{10} \left( \frac{E(ROI_O) - E(ROI_{bg})}{\sigma_{bg}} \right) \quad (18)$$

and the signal-to-noise ratio (SNR) by

$$SNR = 10 \log_{10} \left( \frac{\sum_i (J_{out\,i} - E(ROI_{bg}))}{\sigma_{bg}} \right) \quad (19)$$

where $ROI_O$ and $ROI_{bg}$ are the specific regions of interest within the object and background, respectively; the $\sigma_{bg}$ is the noise in the background region; $E(ROI_{bg})$ is the average signal in the background region of interest (ROI); $(J_{out\,i} - E(ROI_{bg}))$ is the signal at each pixel $i$ in the denoised image.

Furthermore, qualitative evaluation was performed using the overall image quality, contrast, resolution, and noise level. A radiologist with more than 20 years of experience in US imaging blindly rated the outcomes of HCTSpeckle, the most recent method, and the original data. The images were rated on an integer scale from 1 to 5, where higher values indicate better quality. A rating of 5 reflects a significant preference for the resultant image over the original, while a rating of 4 shows a slight preference for the result. A rating of 3 signifies that the two images are equivalent. Conversely, lower ratings (1 and 2) indicate a preference for the original image, with 1 indicating a significant preference and 2 indicating a slight preference. Additionally, alveolar bone segmentation was conducted on dental US data with and without denoising by HCTSpeckle, and the segmentation performance was evaluated using segmentation metrics such as Dice score, precision, specificity, and sensitivity.

## V. RESULTS AND DISCUSSION

We evaluated the assessment of the HCTSpeckle technique along with several denoising methods developed within the last five years, including the four traditional techniques [7], [15], [67], [68] and six DL-based techniques [20], [21], [22], [23], [24], [69].

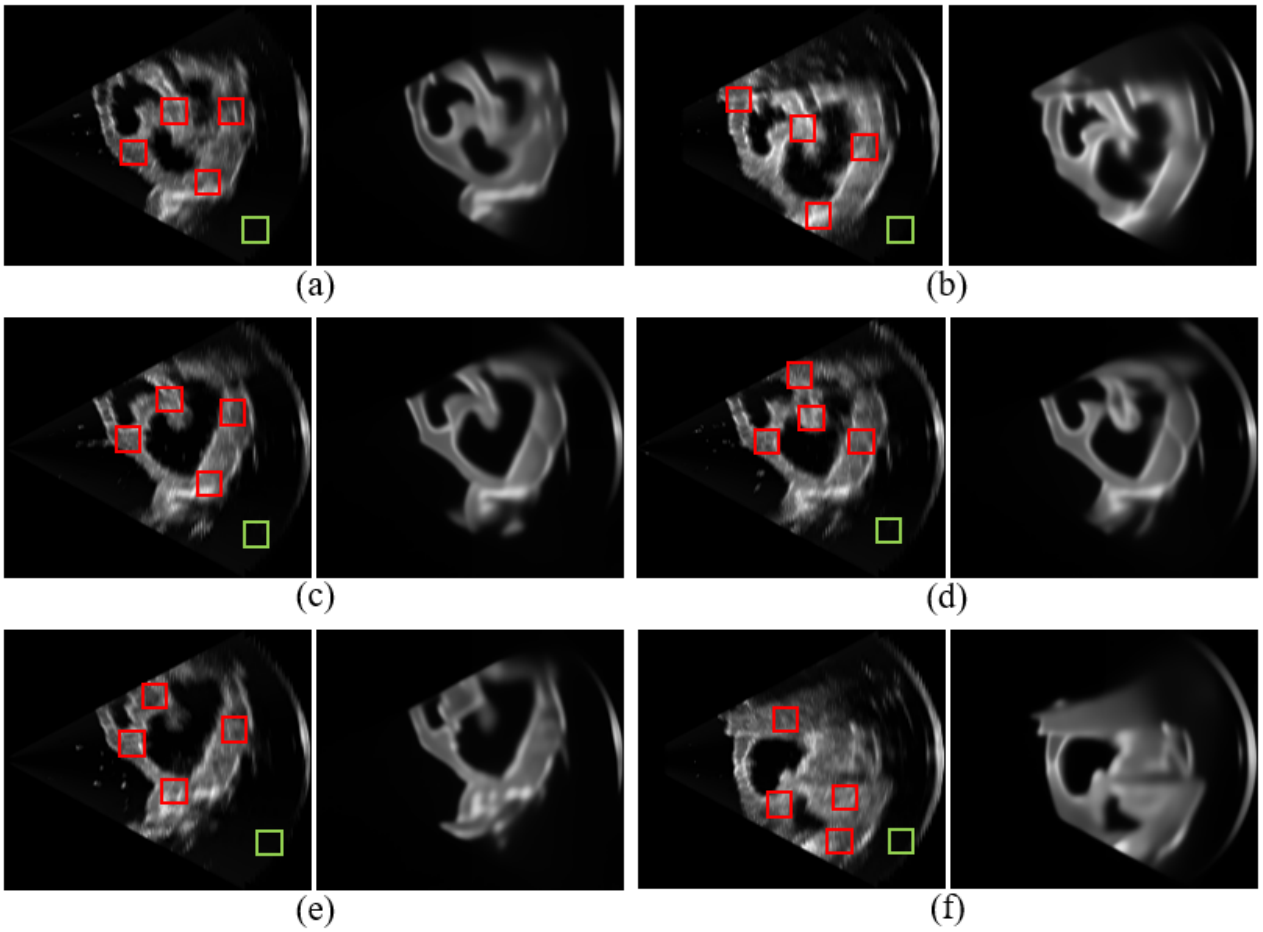



**FIGURE 4.** Qualitative comparison of the denoised results in six real US images. For each set, ROI is 64 × 64: original noisy image (Left), denoised result by the HCTSpeckle (Right). Red ROIs are depicted as foreground regions, and green ROI is depicted as background region. The real US images with four ROIs and background ROIs are marked for computing CNR & SNR.

### A. QUANTITATIVE COMPARISON

The effectiveness of HCTSpeckle was validated using a testing set of fetal head US real images and non-reference metrics such as ISNR, SI, Entropy, and NIQE. The results were presented in Table 4. As illustrated, HCTSpeckle proved superior in effectively reducing noise in US images, achieving remarkable ISNR, SI, Entropy, and NIQE scores of 32.61 ± 11.19, 0.04 ± 0.01, 7.26 ± 2.62, and 4.61 ± 1.23, respectively, and outperformed the existing approaches.

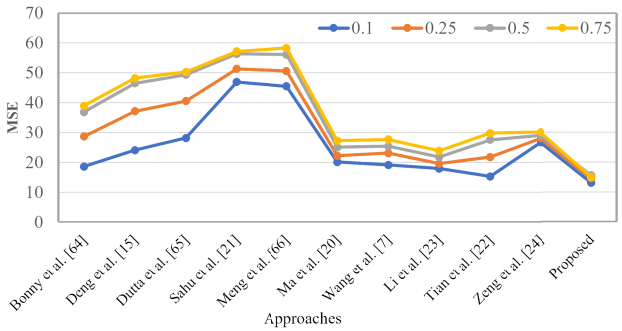
Furthermore, the effectiveness of the denoised results was compared using CNR and SNR metrics on the heart phantom data. Four sets of ROIs were chosen with the foreground regions marked in red and the background marked in green, as shown in Fig. 4. The tabulated values for the average CNR and average SNR are listed in Table 5. Regarding SNR values, the highest average values in the denoised outcome were achieved by HCTSpeckle compared with the original image, demonstrating that noise removal was effectively performed by HCTSpeckle. Slightly higher CNR values were obtained by HCTSpeckle compared with the original values. Overall, the CNR and SNR comparisons indicated that speckle noise was effectively removed by HCTSpeckle while maintaining the integrity of the image details.

**TABLE 4.** Performance comparison of real images of the HC18 testing dataset using average scores of SI, ISNR, Entropy and NIQE. The lower the SI and NIQE values and the higher the ISNR and entropy values, the better the performance. Tr and DL denote the traditional and the DL-based denoising techniques, respectively. The corresponding indices display their best values in bold.

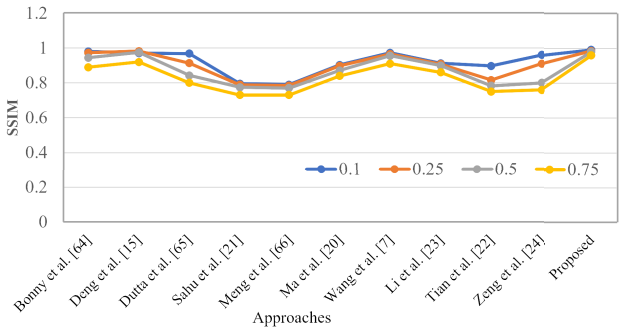
| Approaches | Type | ISNR | SI | Entropy | NIQE |
|---|---|---|---|---|---|
| Sahu et al. [21] | DL | 14.27 ± 9.92 | 0.07 ± 0.06 | 6.25 ± 4.46 | 7.51 ± 2.24 |
| Tian et al. [22] | DL | 14.15 ± 10.08 | 0.07 ± 0.08 | 6.50 ± 3.28 | 6.44 ± 4.13 |
| Meng et al. [69] | DL | 26.85 ± 12.56 | 0.05 ± 0.02 | 6.32 ± 4.33 | 6.19 ± 3.04 |
| Wang et al. [7] | Tr | 24.20 ± 11.68 | 0.05 ± 0.04 | 6.08 ± 2.51 | 6.17 ± 4.15 |
| Zeng et al. [24] | DL | 26.75 ± 10.03 | 0.06 ± 0.04 | 6.30 ± 4.81 | 6.16 ± 4.39 |
| Deng et al. [15] | Tr | 28.83 ± 9.60 | 0.05 ± 0.01 | 6.08 ± 2.09 | 5.92 ± 1.76 |
| Bonny et al. [67] | Tr | 27.69 ± 8.37 | 0.05 ± 0.02 | 6.53 ± 3.29 | 5.83 ± 4.06 |
| Ma et al. [20] | DL | 28.77 ± 10.44 | 0.05 ± 0.04 | 6.09 ± 3.57 | 5.54 ± 4.18 |
| Dutta et al. [68] | Tr | 29.42 ± 9.65 | 0.05 ± 0.03 | 6.19 ± 4.81 | 5.46 ± 3.16 |
| Li et al. [23] | DL | 31.41 ± 10.17 | 0.04 ± 0.03 | 6.83 ± 3.11 | 5.02 ± 1.92 |
| **Proposed** | DL | **32.61 ± 11.19** | **0.04 ± 0.01** | **7.26 ± 2.62** | **4.61 ± 1.23** |

**TABLE 5.** The CNR and SNR comparisons for the original and denoised results by the HCTSpeckle were performed in the six US heart phantom images as shown in Fig. 4. The higher CNR and SNR values denote the best performance. The best scores are bolded.

| Image | CNR (dB) | | SNR (dB) | |
|---|---|---|---|---|
| | Noisy | Denoised | Noisy | Denoised |
| (a) | 1.38 ± 0.76 | **1.82 ± 0.30** | 15.49 ± 1.34 | **22.65 ± 1.39** |
| (b) | 1.05 ± 0.24 | **1.76 ± 0.25** | 14.07 ± 1.17 | **17.90 ± 1.11** |
| (c) | 1.22 ± 0.75 | **1.48 ± 0.77** | 11.37 ± 1.49 | **13.58 ± 2.54** |
| (d) | 1.43 ± 0.17 | **1.98 ± 0.18** | 13.91 ± 0.68 | **18.63 ± 0.61** |
| (e) | 1.06 ± 0.13 | **1.85 ± 0.08** | 10.98 ± 0.42 | **17.53 ± 0.97** |
| (f) | 1.26 ± 0.09 | **1.70 ± 0.16** | 11.01 ± 0.69 | **17.64 ± 1.43** |



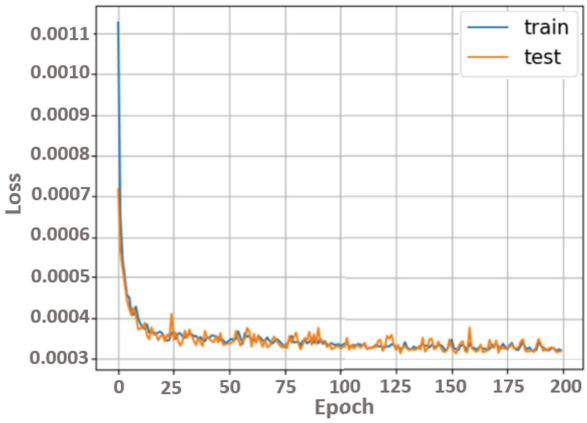

(a) MSE comparison (the lowest is the best)



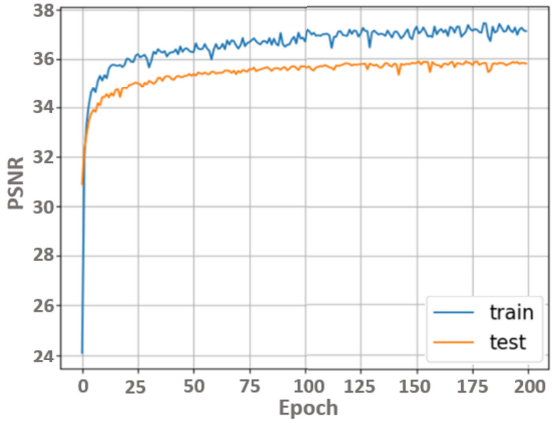

(b) SSIM comparison (highest is best)

**FIGURE 5.** Comparison of different metrics on the BUSI dataset with different noise levels.

In addition, a quantitative evaluation was conducted with different noise levels using MSE, PSNR, and SSIM on the BUSI and HC18 testing sets. The results are plotted in Fig. 5 and summarized in Table 6, respectively. The results demonstrated the strong performance of HCTSpeckle in terms of MSE, PSNR, and SSIM at each noise level on the HC18 dataset. HCTSpeckle achieved a minimum MSE of 13.04 ± 8.27, and a maximum of 15.01 ± 8.74. Additionally, HCTSpeckle excelled in preserving structural similarity, as evident from the SSIM values presented in Table 6. Notably, HCTSpeckle attained high SSIM values of 0.98 ± 0.26 at a noise level of 0.1. Even at a high noise variance of 0.75, the approach still achieved an SSIM value of 0.95 ± 0.31, while some other methods displayed notably poorer performance at this noise level. The obtained results showed that HCTSpeckle excels in preserving intricate details and upholding image quality, which is a critical factor for accurate disease diagnosis by radiologists. Additionally, Fig. 6 presents the variations of loss and PSNR values during the training and testing phases.




(a) Training and Testing losses




(b) PSNR measures

**FIGURE 6.** Training and testing losses, and PSNR measures with respect to epochs for the HC18 dataset.

### B. QUALITATIVE COMPARISON

Fig. 7 presents a visual comparison of the results from the ten most recent existing methods along with HCTSpeckle. Additionally, an enlarged view of the denoised outcomes using recent despeckling approaches such as RED-MAM [23] and REDSENet [24] on dental US data is illustrated in Fig. 8. The noise is clearly visible in Fig. 8(a). In Fig. 8(c), the image exhibits noticeable noise and displays an overly smooth texture, whereas in Fig. 8(b), the texture is sharper than that in Fig. 8(c). Notably, Fig. 8(d) displays sharper details than those in Figs. 8(b) and (c). Furthermore, the enlarged area in Fig. 8(d) exhibits more distinct boundaries than the other denoised results.

Fig. 9 shows a visual comparison of the outcomes and pixel intensity profiles along the highlighted line for the original and denoised results of several CNN-based techniques

**TABLE 6.** Comparisons with the existing approaches on the HC18 dataset at different noise levels ($\sigma$) using average scores of MSE and SSIM. The lower the MSE value and the higher the SSIM values are, the better is the performance. The Tr and DL denote traditional and DL-based denoising techniques, respectively. The corresponding indices display their best values in bold.

| Methods | Type | $\sigma = 0.1$ | | $\sigma = 0.25$ | | $\sigma = 0.5$ | | $\sigma = 0.75$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSE | SSIM | MSE | SSIM | MSE | SSIM | MSE | SSIM |
| Wang et al. [7] | Tr | 25.27 ± 10.53 | 0.72 ± 0.29 | 26.13 ± 9.84 | 0.68 ± 0.21 | 28.55 ± 12.37 | 0.66 ± 0.30 | 30.24 ± 10.43 | 0.63 ± 0.29 |
| Tian et al. [22] | DL | 25.48 ± 11.42 | 0.79 ± 0.61 | 27.64 ± 10.95 | 0.78 ± 0.36 | 29.14 ± 10.37 | 0.77 ± 0.24 | 30.30 ± 11.82 | 0.73 ± 0.29 |
| Sahu et al. [21] | DL | 26.95 ± 10.10 | 0.80 ± 0.43 | 28.32 ± 11.27 | 0.79 ± 0.26 | 29.36 ± 11.53 | 0.77 ± 0.31 | 30.18 ± 11.69 | 0.73 ± 0.48 |
| Ma et al. [20] | DL | 17.65 ± 8.23 | 0.88 ± 0.51 | 18.27 ± 9.84 | 0.87 ± 0.19 | 19.62 ± 8.66 | 0.86 ± 0.34 | 21.36 ± 10.28 | 0.85 ± 0.40 |
| Meng et al. [69] | DL | 19.71 ± 9.97 | 0.90 ± 0.16 | 21.18 ± 14.45 | 0.88 ± 0.24 | 23.69 ± 10.81 | 0.87 ± 0.52 | 24.76 ± 10.62 | 0.86 ± 0.49 |
| Deng et al. [15] | Tr | 24.12 ± 10.71 | 0.95 ± 0.18 | 30.16 ± 11.26 | 0.93 ± 0.31 | 33.48 ± 12.47 | 0.92 ± 0.64 | 37.19 ± 12.52 | 0.90 ± 0.39 |
| Dutta et al. [68] | Tr | 28.20 ± 13.47 | 0.95 ± 0.36 | 31.56 ± 10.24 | 0.91 ± 0.60 | 35.34 ± 11.36 | 0.84 ± 0.53 | 39.27 ± 10.60 | 0.80 ± 0.63 |
| Bonny et al. [67] | Tr | 19.16 ± 10.76 | 0.96 ± 0.31 | 23.13 ± 9.69 | 0.95 ± 0.49 | 25.46 ± 8.57 | 0.94 ± 0.26 | 27.70 ± 10.62 | 0.91 ± 0.37 |
| Zeng et al. [24] | Tr | 14.74 ± 9.60 | 0.96 ± 0.35 | 15.15 ± 8.90 | 0.91 ± 0.56 | 16.07 ± 10.23 | 0.87 ± 0.76 | 16.75 ± 9.38 | 0.87 ± 0.18 |
| Li et al. [23] | DL | 14.98 ± 10.15 | 0.97 ± 0.28 | **14.61 ± 9.23** | **0.97 ± 0.76** | 15.84 ± 10.30 | 0.95 ± 0.33 | 16.91 ± 9.06 | 0.94 ± 0.13 |
| **Proposed** | DL | **13.04 ± 8.27** | **0.98 ± 0.26** | 14.92 ± 9.36 | 0.96 ± 0.33 | **15.43 ± 9.72** | **0.96 ± 0.38** | **15.01 ± 8.74** | **0.95 ± 0.31** |

implemented on a heart phantom US image. From Fig. 9, it is evident that HCTSpeckle generated smoother background regions and better preserved structural information compared to the other DL-based methods such as Sahu et al. [21] and Tian et al. [22], particularly within the selected green and blue ROIs. In addition to the enlarged view and the profile intensity plots, we generated a color plot for the real images and the corresponding denoised outcomes to demonstrate the preservation of structural details such as the boundary, margins and tips of anatomies while removing noise from US images, as shown in Fig. 10.

### C. COMPUTATIONAL EFFICIENCY COMPARISON

In addition to quantitative and qualitative comparisons, the computational complexity of the proposed HCTSpeckle approach was evaluated against the other state-of-the-art DL-based methods from the past five years. Metrics such as trainable parameter count, number of floating-point operations (FLOPs), average inference time for a single image, and average SSIM index were used to assess the computational efficiency.

The numerical results were listed in Table 7, demonstrating that the proposed approach outperformed the other methods in terms of gigaFLOPs, inference time, and denoising performance. While the capsule network-based method had fewer million parameters, the proposed HCTSpeckle excelled with lower gigaFLOPs, shorter inference time, and superior denoising performance. Specifically, HCTSpeckle showcased remarkable computational efficiency with only 17.94 million trainable parameters, 8.90 gigaFLOPs, and an inference time of 82 milliseconds, respectively. It was notably faster than other methods, which exhibited higher parameter counts, gigaFLOPs, and inference times. Moreover, HCTSpeckle achieved the highest average SSIM of $0.96 \pm 0.38$, indicating superior image quality preservation. These results highlighted the efficiency and effectiveness of the proposed HCTSpeckle approach in both computational and qualitative aspects.

### D. COMPARISON WITH SEGMENTATION PERFORMANCE

For the segmentation of alveolar bone in dental US images, a well-established U-Net segmentation network [50] was employed as the segmentation model. U-Net primarily comprised an encoder for capturing image features and a decoder for constructing and localizing segmentation labels. Sampling paths utilized a concatenation operator instead of a sum, and skip connections were designed to convey local information to global information during the up-sampling process.

In the image segmentation phase, the model underwent training for 500 epochs with a batch size of eight, utilizing the Adam optimizer with a learning rate of $10^{-4}$, and the loss function by the Dice coefficient. The U-Net model was trained using 1264 dental US images, with 316 images for testing purposes. All input images and the corresponding ground truth masks were resized to $256 \times 256$. Training data were employed to compute the parameters of the neural networks, which were updated iteratively to minimize the cost function. The best model was used to segment both real and denoised images. The performance of segmenting dental structures in each testing image was evaluated using the Dice coefficient, precision, specificity, and sensitivity. The average results of the segmentation metrics for US images, both with and without denoising performed by HCTSpeckle, were presented in Table 8. The results with the denoising approach achieved higher values for all metrics compared with the segmented outcome of the original images. In summary, the obtained results demonstrated that HCTSpeckle improved the performance of other imaging tasks, such as segmentation. The accurate segmentation of alveolar bone in dental images was crucial for periodontal diagnosis [26].

### E. STATISTICAL COMPARISON

From the testing set of dental US data, 250 images were randomly selected and blindly rated by the radiologist using the above-mentioned scale for the original image and the outcomes of RED-MAM, REDSENet, and HCTSpeckle.

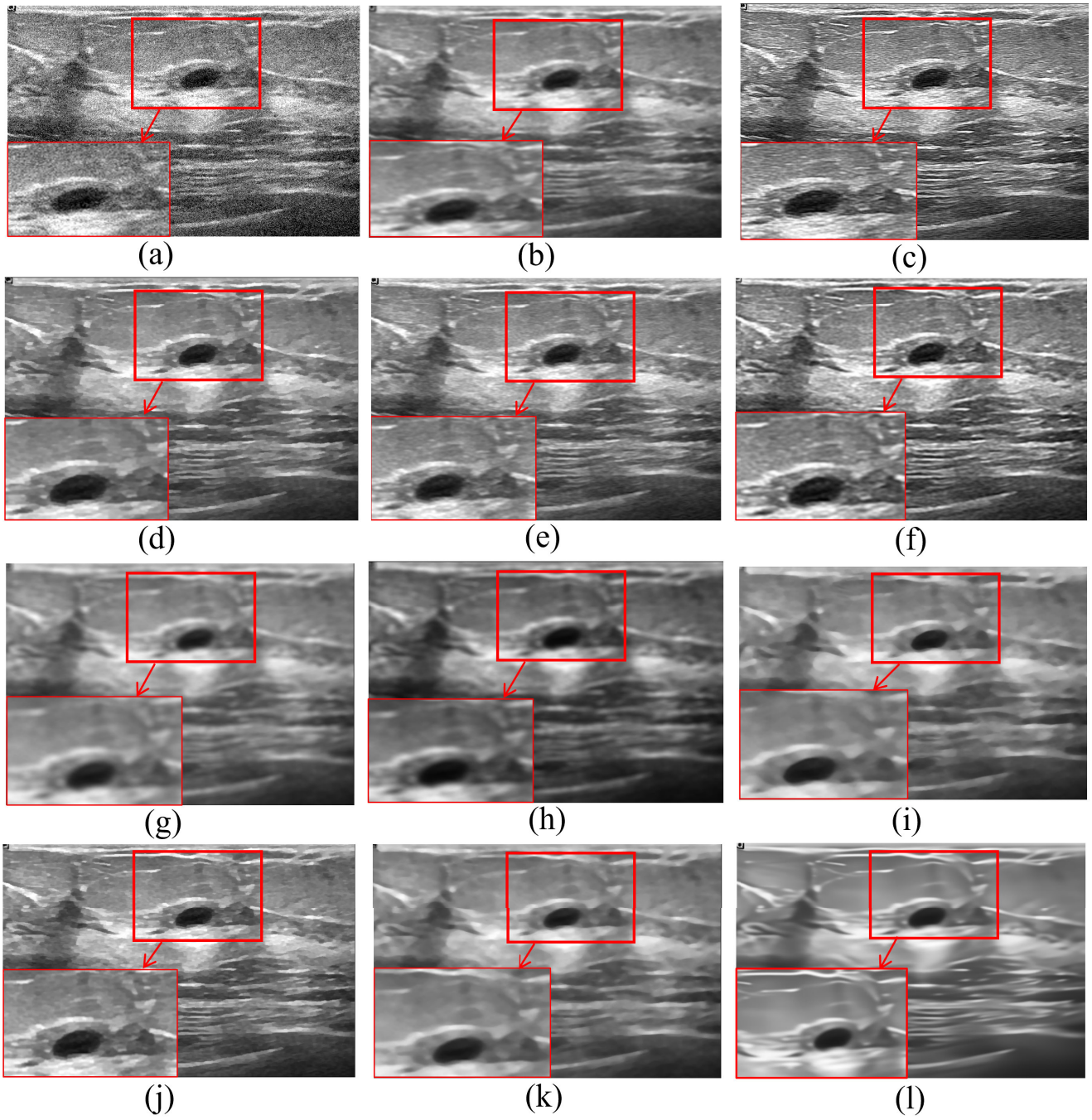



**FIGURE 7.** Visual comparison of the ten top-performing methods on breast US image from the BUSI dataset: (a) Noisy, (b) Ma et al. [20], (c) Sahu et al. [21], (d) Tian et al. [22], (e) Wang et al. [69], (f) Deng et al. [15], (g) Zeng et al. [24], (h) Wang et al. [7], (i) Dutta et al. [68], (j) Li et al. [23], (k) Bonny et al. [67], and (l) HCTSpeckle. The rectangular enclosures indicate the zoomed areas.

The mean outcomes of the assessments conducted by the radiologist are presented in Table 9. The HCTSpeckle achieved values in quality, noise, contrast, and resolution of $4.05 \pm 0.32$, $4.06 \pm 0.37$, $4.05 \pm 0.36$, and $4.04 \pm 0.34$, respectively. It was observed that in terms of image quality, noise, contrast, and resolution, HCTSpeckle's outcomes exhibited a significant advantage compared to the other recent approaches.

HCTSpeckle was compared with RED-MAM and RED-SENet. The experimental results revealed that HCTSpeckle and RED-MAM exhibited significant improvements in image quality, sharpness, and artifacts. HCTSpeckle achieved a higher value for each quality metric than RED-MAM and REDSENet. When comparing the "image quality," the mean differences between HCTSpeckle and RED-MAM and between HCTSpeckle and REDSENet were $-1.128$ and

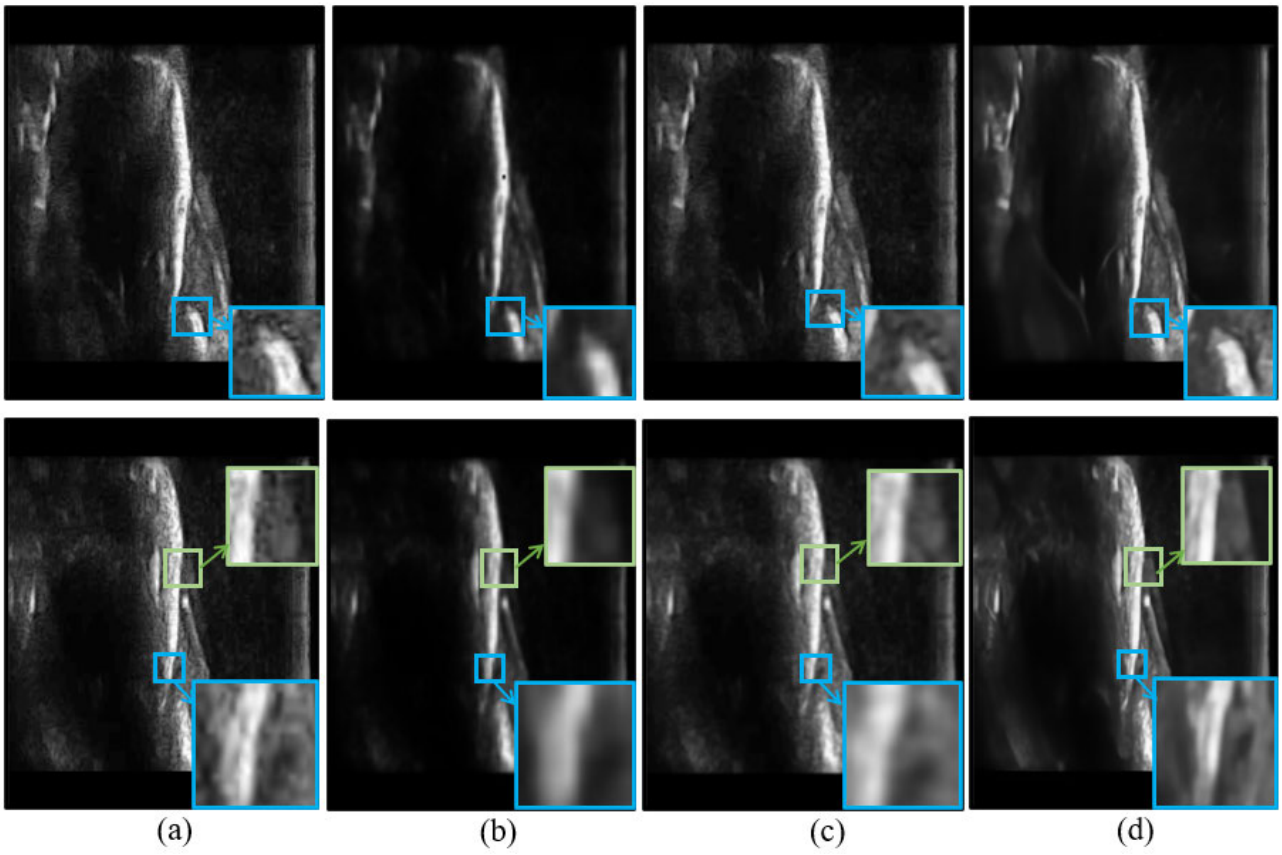



**FIGURE 8.** Enlarged view of the denoised outcomes of the most recent despeckling approaches on the dental US data: (a) Original, (b) Li et al. [23], (c) Zeng et al. [24], and (d) HCTSpeckle. The proposed result (d) exhibits more distinct boundaries & sharper details like alveolar bone crest, gingiva, enamel margins compared to other denoised results (b) and (c).

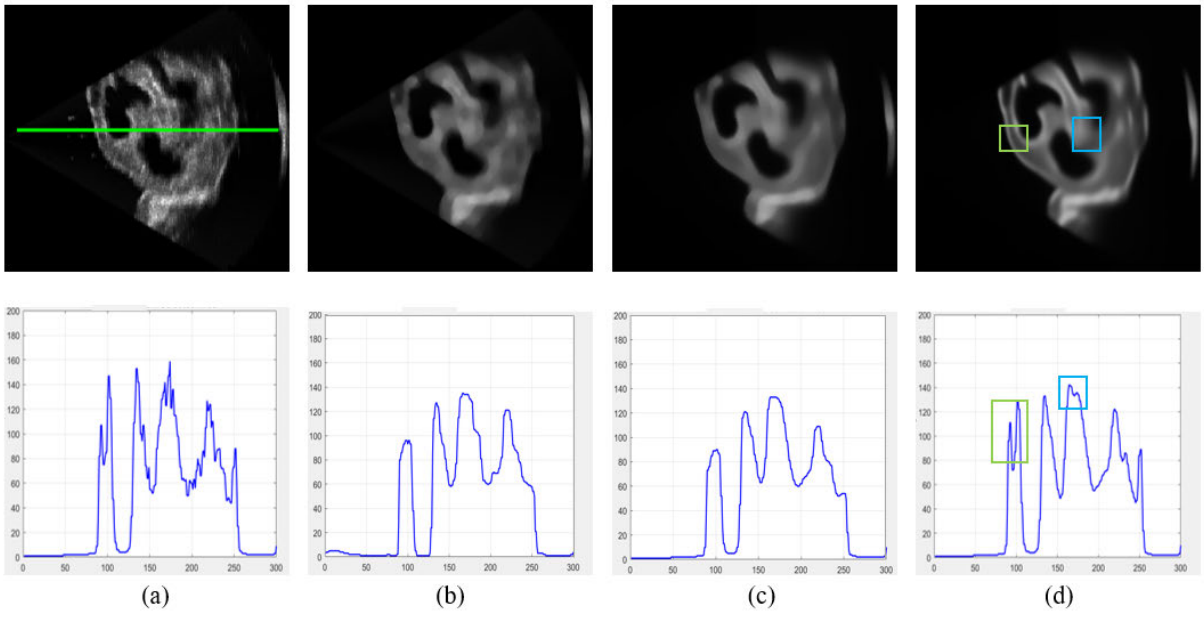



**FIGURE 9.** Visual comparison of outcome and profiles of pixel intensities along the highlighted line for the original and denoised outcomes performed on a heart phantom US image: (a) The heart phantom US image, (b) Sahu et al. [21], (c)Tian et al. [22], and (d) HCTSpeckle.

−2.036, respectively. In "noise level" comparison, the mean differences between HCTSpeckle and RED-MAM and between HCTSpeckle and REDSENet were −1.908 and −2.112, respectively.

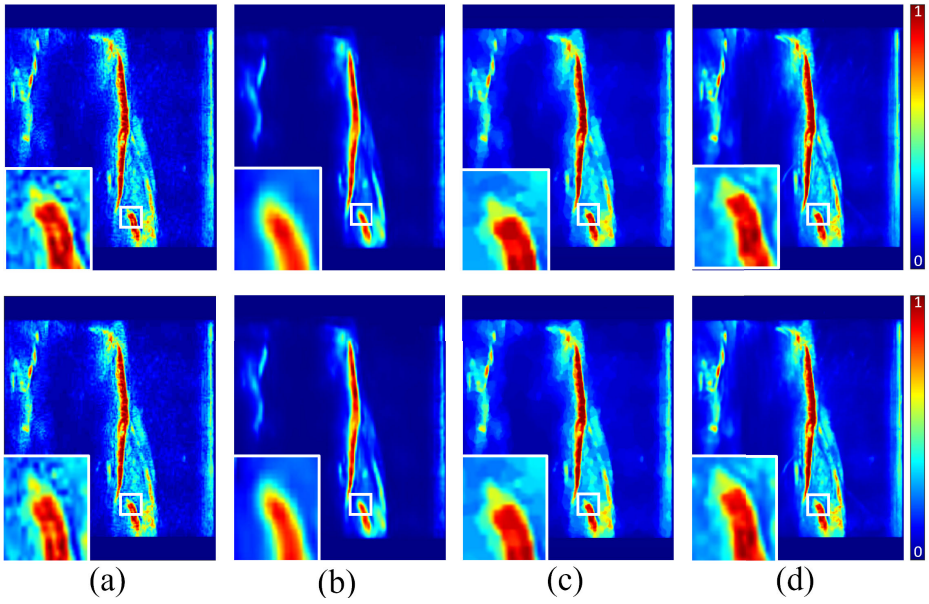



**FIGURE 10.** Qualitative results of the most recent despeckling approaches on the dental US data: (a) Original, (b) Li et al. [23], (c) Sahu et al. [21], and (d) HCTSpeckle. It is essential to preserve the boundary structural details while removing noise. The color plot shows the boundary details of an image. The proposed result (d) exhibits more distinct boundaries & sharper details compared to the other denoised results (b) and (c).

Moreover, for the ''contrast,'' HCTSpeckle consistently had higher means than RED-MAM and REDSENet in both pair comparisons. In the first comparison, the mean difference between HCTSpeckle and RED-MAM was −1.148. In the second comparison, HCTSpeckle also showed a higher mean than REDSENet, with a mean difference of -2.020. Finally, when comparing the ''resolution,'' the mean difference for HCTSpeckle and RED-MAM was −1.168. For the second pair, HCTSpeckle and REDSENet, with a mean difference of −2.004. These results demonstrate that HCTSpeckle consistently outperformed RED-MAM and REDSENet in terms of the mean scores of the evaluated qualitative metrics. The mean differences were statistically significant with a $p$-value of less than 0.001.

### F. ABLATION STUDY

The effectiveness of the proposed denoising technique was primarily based on the SR block. Consequently, we compared the denoising and model reconstruction performances using the HC18 dataset to assess the requirements for our SR block. A diagram illustrating the block insertion point is shown in Fig. 11(a), and the corresponding experimental results are presented in Fig. 12. Incorporating the SR block in positions one and two within an encoder-decoder structure results in greater improvement than positioning the SR block in one or two individually.

As shown in Fig. 13, quantitative evaluations of each component within the HCTSpeckle model were performed individually. These experiments followed the findings outlined in Fig. 12, and the subsequent experiments were conducted at positions 1 and 2. The results revealed that both the individual swin and res blocks significantly impact the model's performance, with the swin block exhibiting a more pronounced effect than the res block. Furthermore, this study explored the collective effect of incorporating SR blocks into the denoising model. Fig. 2 illustrates the structure of an encoder-decoder (U-shaped network), with ''U + Swin'' denoting the inclusion of swin blocks in the U-shaped network, as depicted in Fig. 11(b). ''U + Swin + Res'' signifies the incorporation of res blocks, whereas ''Swin + Res'' is a sequential combination of an SR block (U+seq) (Fig. 11(c)). Lastly, ''U+ Swin & Res'' represents the parallel integration of the swin block and the res block (U+par), as illustrated in Fig. 11(d). The experimental results indicated that combining blocks in parallel, particularly within the HCTSpeckle model, led to a more noticeable enhancement in the denoising model as shown in Fig. 13.

Furthermore, adding ISBs to the denoiser led to further significant improvements in MSE, PSNR and SSIM. Specifically, the best performance was achieved with the inclusion of the ISB to the parallel combination (U+par+ISB), yielding the lowest MSE of 13.04, the highest PSNR of 37.85, and the highest SSIM of 0.98. These results highlighted the effectiveness of each component and their synergistic impact when combined, demonstrating the robustness and superior performance of the proposed HCTSpeckle denoiser.

### G. DISCUSSION

HCTspeckle was proposed as a hybrid network that combines CNNs and Transformers. Two notable hybrid

**TABLE 7.** Computational efficiency comparison of the proposed approach with the state-of-the-art DL-based methods. The abbreviations are denoted as M for Million, FLOPs for number of floating point operations, and ms for milliseconds. The best performance values are bolded.

| Method | No. of Trainable Parameters (M) | FLOPs (G) | Inference Time (ms) | Average SSIM |
|---|---|---|---|---|
| Tian et al. [22] | 137.54 | 10.62 | 301 | 0.77 ± 0.24 |
| Sahu et al. [21] | 167.38 | 12.50 | 377 | 0.77 ± 0.31 |
| Ma et al. [20] | 264.82 | 11.51 | 537 | 0.86 ± 0.34 |
| Meng et al. [69] | **6.14** | 8.94 | 95 | 0.87 ± 0.52 |
| Zeng et al. [24] | 185.73 | 10.46 | 752 | 0.87 ± 0.76 |
| Li et al. [23] | 210.3 | 21.47 | 91 | 0.95 ± 0.33 |
| **Proposed** | 17.94 | **8.90** | **82** | **0.96 ± 0.38** |

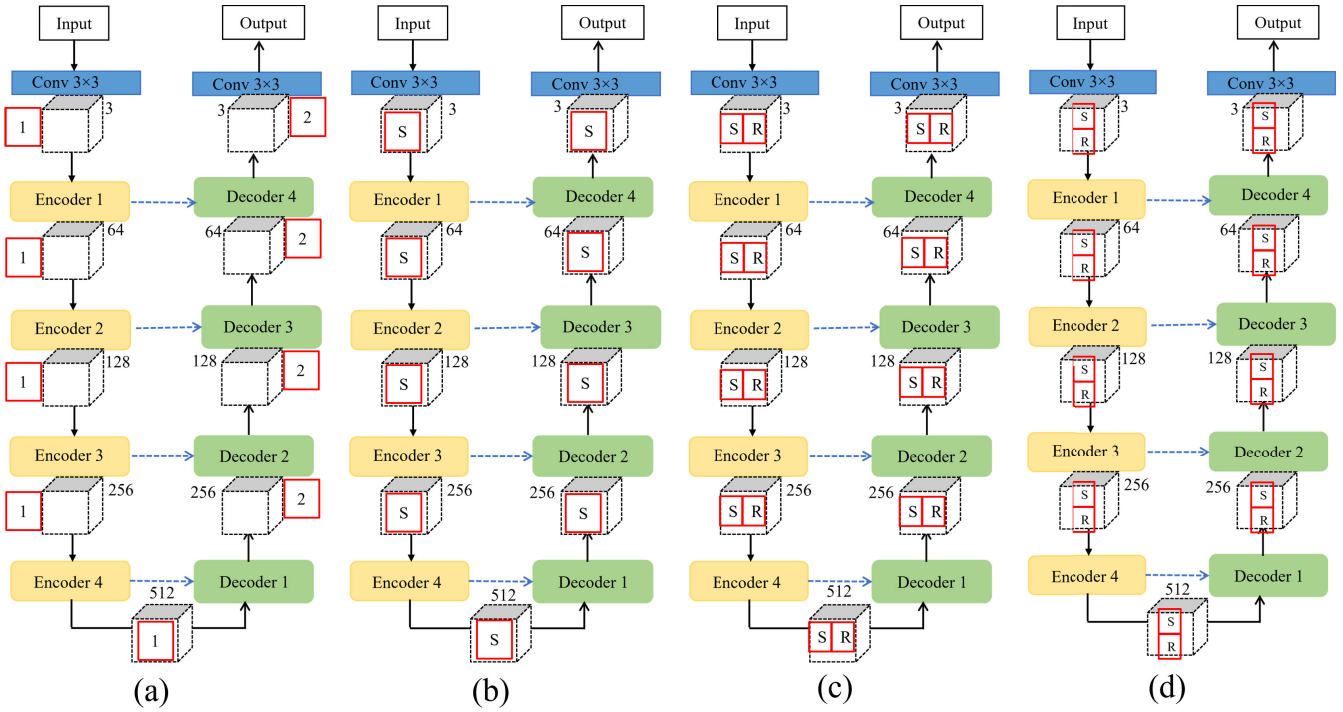



**FIGURE 11.** The block diagrams illustrating the effectiveness of the proposed HCTSpeckle method. (a) The position where the fusion SR block is inserted. (b) A U-shaped network with swin blocks. (c) A U-shaped network with a sequence of swin and res blocks. (d) A U-shaped network with a parallel connection of swin and res blocks.

**TABLE 8.** Quantity assessment of segmentation performance using a U-Net model in dental US data.

| Metrics (%) | w/o denoising | with denoising |
|---|---|---|
| Dice score | 62.48 ± 5.10 | **71.35 ± 6.41** |
| Precision | 70.32 ± 6.24 | **76.98 ± 8.80** |
| Specificity | 80.63 ± 2.58 | **88.36 ± 1.24** |
| Sensitivity | 69.17 ± 4.79 | **76.60 ± 7.11** |

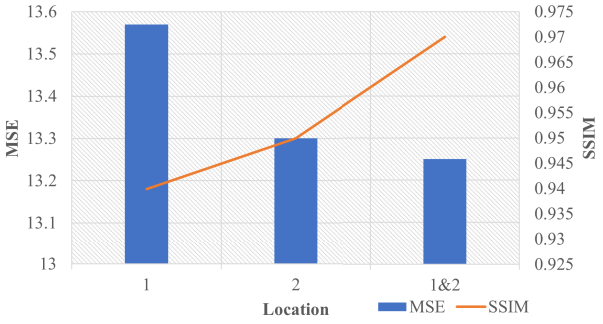



**FIGURE 12.** The fusion (SR) blocks are inserted at different positions under the same noise level on HC18 dataset.

CNN-Transformer denoising networks have been proposed for real-world image denoising applications. The first uses a swin transformer in the encoder section and a single convolution in the decoder section, as detailed in [70]. This design aims to reduce the computational complexity and achieve a balance between the model capacity and computational cost for real-world scenes; however, it limits

**TABLE 9.** The image evaluation ratings obtained by an experienced radiologist on 250 dental US images on a scale of 1 to 5; the greater the value, the higher the image quality.

| Approach | Image Quality | Noise | Contrast | Resolution |
|---|---|---|---|---|
| REDSENet [24] | 2.01 ± 0.48 | 1.95 ± 0.39 | 2.03 ± 0.43 | 2.04 ± 0.49 |
| RED-MAM [23] | 2.92 ± 0.52 | 2.15 ± 0.45 | 2.90 ± 0.55 | 2.88 ± 0.56 |
| **HCTSpeckle** | **4.05 ± 0.32** | **4.06 ± 0.37** | **4.05 ± 0.36** | **4.04 ± 0.34** |

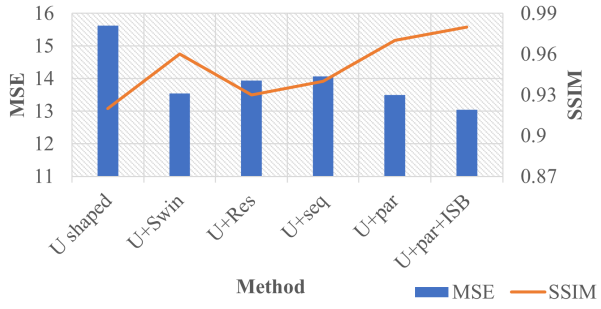



**FIGURE 13.** Effectiveness of each component of the proposed denoiser under the same noise level.

the denoising performance by over smoothing effects for other noise characteristics and other applications. The second network, TC-net [71], reduces the image noise by incorporating transformers and convolutions. It adjusts the feature size using an input adjustment module and effectively extracts the low-level image features. Although TC-net achieves a better denoising performance, it produces poor visual quality for images with other noise variants.

In medical imaging, the only one hybrid network, which is proposed for low-dose CT images is the HCformer [72], which has not been applied to US images. The HCformer uses CNNs for initial feature extraction and the final image restoration, whereas the encoder and decoder blocks are built solely using transformer architectures. Despite its good feature extraction and noise removal capabilities, the HCformer produces over smoothed images and requires significant computational resources.

However, there are four major differences between our proposed HCTspeckle and the recent hybrid CNN-transformer approaches. Firstly, HCTspeckle is the first hybrid CNN-Transformer-based denoising network built specifically for US imaging. HCTspeckle focuses on combining a transformer variant with U-Net. Secondly, the proposed HCTspeckle primarily employs a swin transformer block to reduce computational cost and capture global interactive features while using CNNs to extract the local modeling features. Thirdly, HCTSpeckle introduces a new fusion block that integrates the local modeling capability of res blocks and the non-local modeling ability of swin transformer block through $1 \times 1$ convolution. In contrast, existing approaches use a transformer block with depth-wise convolutional layers. Finally, the improved swin block is introduced at the end of the proposed encoder-decoder network to improve the robustness of HCTSpeckle by extracting more significant features, thereby enhancing the noise removal effectiveness.

## VI. CONCLUSION

In this paper, a novel hybrid CNN-transformer network (HCTSpeckle) is proposed for ultrasound image denoising. The HCTSpeckle network integrates the strengths of the CNN blocks for local modeling and swin transformer blocks for non-local modeling. HCTSpeckle is trained on various noise levels to effectively handle complex noise in the US images. Its effectiveness has been validated using four distinct US datasets. Quantitative results has demonstrated that HCT-Speckle outperforms the other compared methods in terms of evaluation metrics such as MSE, SSIM, CNR and SNR. The visual comparisons has also demonstrated that HCTSpeckle outperforms in noise reduction and structure preservation. Furthermore, using HCTSpeckle as a preprocessing step for alveolar bone segmentation significantly improved the segmentation accuracy. Future work will focus on reducing the computational complexity without compromising the performance of the model, so that the proposed algorithm can be applied in clinical work.

## ACKNOWLEDGMENT

The authors would like to acknowledge the aforementioned funding organizations to make this study possible. Lawrence H. Le and Kumaradevan Punithakumar shared senior authorship.

## REFERENCES

[1] H. Yu, M. Ding, X. Zhang, and J. Wu, "PCANet based nonlocal means method for speckle noise removal in ultrasound images," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0205390, doi: 10.1371/journal.pone.0205390.

[2] Z. Tao, H. D. Tagare, and J. D. Beaty, "Evaluation of four probability distribution models for speckle in clinical cardiac ultrasound images," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1483–1491, Nov. 2006, doi: 10.1109/TMI.2006.881376.

[3] P. Coupe, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2221–2229, Oct. 2009, doi: 10.1109/TIP.2009.2024064.

[4] P. Kokil and S. Sudharson, "Despeckling of clinical ultrasound images using deep residual learning," *Comput. Methods Programs Biomed.*, vol. 194, no. 1, Oct. 2020, Art. no. 105477, doi: 10.1016/j.cmpb.2020.105477.

[5] H. Yu, L. Li, M. Zheng, W. Qiu, and M. Ding, "Despecking of ultrasound image using LENet-based nonlocal-means method," in *Medical Imaging: Ultrasonic Imaging and Tomography*, vol. 12932. Bellingham, WA, USA: SPIE, 2024, pp. 276–283, doi: 10.1117/12.3006341.

[6] C. A. N. Santos, D. L. N. Martins, and N. D. A. Mascarenhas, "Ultrasound image despeckling using stochastic distance-based BM3D," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2632–2643, Jun. 2017, doi: 10.1109/TIP.2017.2685339.

[7] S. Wang, T.-Z. Huang, X.-L. Zhao, J.-J. Mei, and J. Huang, "Speckle noise removal in ultrasound images by first- and second-order total variation," *Numer. Algorithms*, vol. 78, no. 2, pp. 513–533, Jun. 2018, doi: 10.1007/s11075-017-0386-x.

[8] A. A. Yahya, J. Tan, B. Su, M. Hu, Y. Wang, K. Liu, and A. N. Hadi, "BM3D image denoising algorithm based on an adaptive filtering," *Multimedia Tools Appl.*, vol. 79, nos. 27–28, pp. 20391–20427, Jul. 2020, doi: 10.1007/s11042-020-08815-8.

[9] N. J. Habeeb, "Medical image denoising with Wiener filter and high boost filtering," *Iraqi J. Sci.*, vol. 64, no. 6, pp. 4023–4035, Jun. 2023, doi: 10.24996/ijs.2023.64.6.40.

[10] P. V. Sudeep, P. Palanisamy, J. Rajan, H. Baradaran, L. Saba, A. Gupta, and J. S. Suri, "Speckle reduction in medical ultrasound images using an unbiased non-local means method," *Biomed. Signal Process. Control*, vol. 28, pp. 1–8, Jul. 2016, doi: 10.1016/j.bspc.2016.03.001.

[11] A. Sharma and B. P. Shrivastava, "Complex wavelet transform with progressive network for medical imaging super resolution," *Multimedia Tools Appl.*, vol. 83, pp. 1–19, May 2024, doi: 10.1007/s11042-024-19448-6.

[12] A. E. Ilesanmi, O. P. Idowu, U. Chaumrattanakul, and S. S. Makhanov, "Multiscale hybrid algorithm for pre-processing of ultrasound images," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102396, doi: 10.1016/j.bspc.2020.102396.

[13] J. Yang, J. Fan, D. Ai, X. Wang, Y. Zheng, S. Tang, and Y. Wang, "Local statistics and non-local mean filter for speckle noise reduction in medical ultrasound image," *Neurocomputing*, vol. 195, pp. 88–95, Jun. 2016, doi: 10.1016/j.neucom.2015.05.140.

[14] E. A. Radhi and M. Y. Kamil, "Anisotropic diffusion method for speckle noise reduction in breast ultrasound images," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 2, pp. 621–631, 2024, doi: 10.22266/ijies2024.0430.50.

[15] L. Deng, H. Zhu, Z. Yang, and Y. Li, "Hessian matrix-based fourth-order anisotropic diffusion filter for image denoising," *Opt. Laser Technol.*, vol. 110, pp. 184–190, Feb. 2019, doi: 10.1016/j.optlastec.2018.08.043.

[16] L. Jain and P. Singh, "A novel wavelet thresholding rule for speckle reduction from ultrasound images," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4461–4471, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.009.

[17] D. Vilimek, J. Kubicek, M. Golian, R. Jaros, R. Kahankova, P. Hanzlikova, D. Barvik, A. Krestanova, M. Penhaker, M. Cerny, O. Prokop, and M. Buzga, "Comparative analysis of wavelet transform filtering systems for noise reduction in ultrasound images," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0270745, doi: 10.1371/journal.pone.0270745.

[18] R. G. Gavaskar and K. N. Chaudhury, "Fast adaptive bilateral filtering," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 779–790, Feb. 2019, doi: 10.1109/TIP.2018.2871597.

[19] O. Karaoğlu, H. Ş. Bilge, and İ. Uluer, "Removal of speckle noises from ultrasound images using five different deep learning networks," *Eng. Sci. Technol., Int. J.*, vol. 29, no. 1, May 2022, Art. no. 101030, doi: 10.1016/j.jestch.2021.06.010.

[20] Y. Ma, F. Yang, and A. Basu, "Edge-guided CNN for denoising images from portable ultrasound devices," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6826–6833, doi: 10.1109/ICPR48806.2021.9412758.

[21] A. Sahu, K. P. S. Rana, and V. Kumar, "An application of deep dual convolutional neural network for enhanced medical image denoising," *Med. Biol. Eng. Comput.*, vol. 61, no. 5, pp. 991–1004, May 2023, doi: 10.1007/s11517-022-02731-9.

[22] C. Tian, Y. Xu, W. Zuo, B. Du, C.-W. Lin, and D. Zhang, "Designing and training of a dual CNN for image denoising," *Knowl. Syst.*, vol. 226, Aug. 2021, Art. no. 106949, doi: 10.1016/j.knosys.2021.106949.

[23] Y. Li, X. Zeng, Q. Dong, and X. Wang, "RED-MAM: A residual encoder-decoder network based on multi-attention fusion for ultrasound image denoising," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104062, doi: 10.2139/ssrn.4081003.

[24] Zeng, L. Xianhua, G. Yancheng, Z. Ge, and Xueting, "Channel adaptive ultrasound image denoising method based on residual encoder-decoder networks," *J. Electron. Inf. Technol.*, vol. 44, no. 7, pp. 2547–2558, 2022, doi: 10.11999/JEIT210331.

[25] P. N. Srinivasu, V. S. Sravya, V. D. Tagore, V. Vinoothna, and Y. Mokshada, "CBCD: Comprehensive analysis for bone cancer diagnosis through MRI imaging," in *Proc. 1st Int. Conf. Cognit., Green Ubiquitous Comput. (IC-CGU)*, Mar. 2024, pp. 1–6, doi: 10.1109/ic-cgu58078.2024.10530830.

[26] K. C. T. Nguyen, D. Q. Duong, F. T. Almeida, P. W. Major, N. R. Kaipatur, T. T. Pham, E. H. M. Lou, M. Noga, K. Punithakumar, and L. H. Le, "Alveolar bone segmentation in intraoral ultrasonographs with machine learning," *J. Dental Res.*, vol. 99, no. 9, pp. 1054–1061, Aug. 2020, doi: 10.1177/0022034520920593.

[27] P. N. Srinivasu, U. Sirisha, K. Sandeep, S. P. Praveen, L. P. Maguluri, and T. Bikku, "An interpretable approach with explainable AI for heart stroke prediction," *Diagnostics*, vol. 14, no. 2, p. 128, Jan. 2024, doi: 10.3390/diagnostics14020128.

[28] A. Sharma and B. P. Shrivastava, "Medical image super-resolution using correlation filter interleaved progressive convolution network (CFIPC)," *Electron. Lett.*, vol. 58, no. 9, pp. 360–362, Apr. 2022, doi: 10.1049/ell2.12467.

[29] A. Saleh Ahmed, W. H. El-Behaidy, and A. A. A. Youssif, "Medical image denoising system based on stacked convolutional autoencoder for enhancing 2-dimensional gel electrophoresis noise reduction," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102842, doi: 10.1016/j.bspc.2021.102842.

[30] S. Cammarasana, P. Nicolardi, and G. Patanè, "Real-time denoising of ultrasound images based on deep learning," *Med. Biol. Eng. Comput.*, vol. 60, no. 8, pp. 2229–2244, Aug. 2022, doi: 10.1007/s11517-022-02573-5.

[31] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound image enhancement using structure oriented adversarial network," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1349–1353, Sep. 2018, doi: 10.1109/LSP.2018.2858147.

[32] H. G. Khor, G. Ning, X. Zhang, and H. Liao, "Ultrasound speckle reduction using wavelet-based generative adversarial network," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3080–3091, Jul. 2022, doi: 10.1109/JBHI.2022.3144628.

[33] L. Zhang and J. Zhang, "Ultrasound image denoising using generative adversarial networks with residual dense connectivity and weighted joint loss," *PeerJ Comput. Sci.*, vol. 8, p. e873, Feb. 2022, doi: 10.7717/peerj-cs.873.

[34] F. Dietrichson, E. Smistad, A. Ostvik, and L. Løvstakken, "Ultrasound speckle reduction using generative adversial networks," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4, doi: 10.1109/ULTSYM.2018.8579764.

[35] J. Liu, C. Li, L. Liu, H. Chen, H. Han, B. Zhang, and Q. Zhang, "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105150, doi: 10.1016/j.bspc.2023.105150.

[36] J. Lee, J. Jeon, Y. Hong, D. Jeong, Y. Jang, B. Jeon, H. J. Baek, E. Cho, H. Shim, and H.-J. Chang, "Generative adversarial network with radiomic feature reproducibility analysis for computed tomography denoising," *Comput. Biol. Med.*, vol. 159, Jun. 2023, Art. no. 106931, doi: 10.1016/j.compbiomed.2023.106931.

[37] A. Sivaanpu, K. Punithakumar, K. Thanikasalam, M. Noga, R. Zheng, D. Ta, E. H. M. Lou, and L. H. Le, "A lightweight ultrasound image denoiser using parallel attention modules and capsule generative adversarial network," *Informat. Med. Unlocked*, vol. 50, Jan. 2024, Art. no. 101569, doi: 10.1016/j.imu.2024.101569.

[38] T. Chen and C. Chefd'Hotel, "Deep learning based automatic immune cell detection for immunohistochemistry images," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2014, pp. 532–540, doi: 10.1007/978-3-319-10581-9_3.

[39] M. Juneja, S. K. Saini, S. Kaul, R. Acharjee, N. Thakur, and P. Jindal, "Denoising of magnetic resonance imaging using Bayes shrinkage based fused wavelet transform and autoencoder based deep learning approach," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102844, doi: 10.1016/j.bspc.2021.102844.

[40] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, no. 1, 2018, Art. no. 7068349, doi: 10.1155/2018/7068349.

[41] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018, doi: 10.1109/TIP.2018.2839891.

[42] J. Zhang, Z. Shangguan, W. Gong, and Y. Cheng, "A novel denoising method for low-dose CT images based on transformer and CNN," *Comput. Biol. Med.*, vol. 163, Sep. 2023, Art. no. 107162, doi: 10.1016/j.compbiomed.2023.107162.

[43] L. Li, X. Yu, Z. Jin, Z. Zhao, X. Zhuang, and Z. Liu, "FDnCNN-based image denoising for multi-labfel localization measurement," *Measurement*, vol. 152, Feb. 2020, Art. no. 107367, doi: 10.1016/j.measurement.2019.107367.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[45] X.-J. Mao, C. Shen, and Y. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. NIPS*, 2016, pp. 532–540.

[46] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017, doi: 10.1109/TMI.2017.2715284.
[47] M. Ran, J. Hu, Y. Chen, H. Chen, H. Sun, J. Zhou, and Y. Zhang, "Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network," *Med. Image Anal.*, vol. 55, pp. 165–180, Jul. 2019, doi: 10.1016/j.media.2019.05.001.
[48] W. Chorney, H. Wang, L. He, S. Lee, and L.-W. Fan, "Convolutional block attention autoencoder for denoising electrocardiograms," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105242, doi: 10.1016/j.bspc.2023.105242.
[49] Y. Lan and X. Zhang, "Real-time ultrasound image despeckling using mixed-attention mechanism based residual UNet," *IEEE Access*, vol. 8, pp. 195327–195340, 2020, doi: 10.1109/ACCESS.2020.3034230.
[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent. (MICCAI)*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 532–540, doi: 10.5555/3295222.3295349.
[52] Z. Zhang, L. Yu, X. Liang, W. Zhao, and L. Xing, "TransCT: Dual-path transformer for low dose computed tomography," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 55–64, doi: 10.1007/978-3-030-87231-1_6.
[53] D. Wang, Z. Wu, and H. Yu, "TED-net: Convolution-free T2T vision transformer-based encoder-decoder dilation network for low-dose CT denoising," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 532–540, doi: 10.1007/978-3-030-87589-3_43.
[54] A. Luthra, H. Sulakhe, T. Mittal, A. Iyer, and S. K. Yadav, "Eformer: Edge enhancement based transformer for medical image denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Sep. 2021, pp. 2667–2677.
[55] J. Zhang, Y. Zhang, J. Gu, J. Dong, L. Kong, and X. Yang, "Xformer: Hybrid X-shaped transformer for image denoising," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024, pp. 1–13.
[56] L. Marcos, P. Babyn, and J. Alirezaie, "Pure vision transformer (CT-ViT) with Noise2Neighbors interpolation for low-dose CT image denoising," *J. Imag. Informat. Med.*, vol. 37, no. 5, pp. 2669–2687, Apr. 2024, doi: 10.1007/s10278-024-01108-8.
[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[58] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17662–17672.
[59] T. L. A. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. V. Ginneken, "Automated measurement of fetal head circumference using 2D ultrasound images," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0200412, doi: 10.1371/journal.pone.0200412.
[60] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863, doi: 10.1016/j.dib.2019.104863.
[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.
[62] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369, doi: 10.1109/ICPR.2010.579.
[63] C. P. Loizou and C. S. Pattichis, *Despeckle Filtering for Ultrasound Imaging and Video*, vol. 1, 2nd ed. Cham, Switzerland: Springer, 2015, pp. 532–540, doi: 10.1007/978-3-031-01523-6.
[64] S. G. Dellepiane and E. Angiati, "Quality assessment of despeckled SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 691–707, Feb. 2014, doi: 10.1109/JSTARS.2013.2279501.
[65] A. M. Reza and U. Techavipoo, "An ultrasound image despeckling approach based on principle component analysis," *Int. J. Image Process.*, vol. 8, no. 4, pp. 156–177, 2014.
[66] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt, J. M. Boone, and E. J. Goldschmidt, "The essential physics of medical imaging," *Int. J. Med. Phys. Res. Practise*, vol. 30, no. 7, p. 1936, Jul. 2003, doi: 10.1118/1.1585033.
[67] S. Bonny, Y. J. Chanu, and K. M. Singh, "Speckle reduction of ultrasound medical images using Bhattacharyya distance in modified non-local mean filter," *Signal, Image Video Process.*, vol. 13, no. 2, pp. 299–305, Mar. 2019, doi: 10.1007/s11760-018-1357-y.
[68] S. Dutta, B. Georgeot, D. Kouamé, D. Garcia, and A. Basarab, "Adaptive contrast enhancement of cardiac ultrasound images using a deep unfolded many-body quantum algorithm," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2022, pp. 1–4, doi: 10.1109/IUS54386.2022.9958691.
[69] M. Wang, W. Zhu, K. Yu, Z. Chen, F. Shi, Y. Zhou, Y. Ma, Y. Peng, D. Bao, S. Feng, L. Ye, D. Xiang, and X. Chen, "Semi-supervised capsule cGAN for speckle noise reduction in retinal OCT images," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1168–1183, Apr. 2021, doi: 10.1109/TMI.2020.3048975.
[70] M. Zhao, G. Cao, X. Huang, and L. Yang, "Hybrid transformer-CNN for real image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1252–1256, 2022, doi: 10.1109/LSP.2022.3176486.
[71] T. Xue and P. Ma, "TC-net: Transformer combined with CNN for image denoising," *Int. J. Speech Technol.*, vol. 53, no. 6, pp. 6753–6762, Mar. 2023, doi: 10.1007/s10489-022-03785-w.
[72] J. Yuan, F. Zhou, Z. Guo, X. Li, and H. Yu, "HCformer: Hybrid CNN-transformer for LDCT image denoising," *J. Digit. Imag.*, vol. 36, no. 5, pp. 2290–2305, Jun. 2023, doi: 10.1007/s10278-023-00842-9.

**ANPARASY SIVAANPU** (Graduate Student Member, IEEE) received the bachelor's degree in computer science from the University of Jaffna, Jaffna, Sri Lanka, in 2020. She is currently pursuing the Ph.D. degree in medical sciences with the University of Alberta, Alberta, Canada, under the supervision of Dr. Lawrence H. Le and Dr. Kumaradevan Punithakumar. Her research interests include medical image analysis, medical image processing, and artificial intelligence.

**KUMARADEVAN PUNITHAKUMAR** (Senior Member, IEEE) received the B.Sc.Eng. degree (Hons.) in electronic and telecommunication engineering from the University of Moratuwa and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from McMaster University. From 2001 to 2002, he was an Instructor with the Department of Electronic and Telecommunication Engineering, University of Moratuwa. In 2008, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, McMaster University. From 2008 to 2012, he was an Imaging Research Scientist with GE Healthcare, Canada. He is currently an Associate Professor and the AHS Chair of diagnostic imaging with the Department of Radiology and Diagnostic Imaging, University of Alberta; and an Operational and Computational Director of the Servier Virtual Cardiac Centre, Mazankowski Alberta Heart Institute. His research interests include medical image analysis and visualization, information fusion, object tracking, and nonlinear filtering. He was a recipient of the Industrial Research and Development Fellowship from the National Sciences and Engineering Research Council of Canada, in 2008.

**RUI ZHENG** (Member, IEEE) received the B.S. and M.S. degrees from the Department of Engineering Physics, Tsinghua University, Beijing, China, in 2000 and 2002, respectively, and the Ph.D. degree in physics and biomedical engineering from the University of Alberta, Edmonton, AB, Canada, in 2011. From 2012 to 2013, she was a Postdoctoral Fellow with the Laboratory of Mechanics and Acoustics, CNRS, Paris, France. From 2013 to 2017, she was a Research Associate with the Department of Surgery, University of Alberta, and the Glenrose Rehabilitation Hospital, Alberta Health Services, Edmonton. She is currently an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. Her research interests include biomedical ultrasound modeling and imaging methods based on deep learning, AI applications on medical ultrasound diagnosis, biomedical ultrasound signal processing, and the development of integrated mobile ultrasound systems.

**MICHELLE NOGA** is currently a Professor with the Department of Radiology and Diagnostic Imaging, University of Alberta, and the Mazankowski Alberta Heart Institute Medical Imaging Consultants Research Chair. She is also a Radiologist in pediatric cardiac imaging. Her research interests include 3D display of medical imaging and automated segmentation of medical imaging.

**DEAN TA** (Senior Member, IEEE) received the M.S. degree from the Institute of Acoustics, Shaanxi Normal University, Xi'an, China, in 1999, and the Ph.D. degree from Tongji University, Shanghai, China, in 2002. He was a Postdoctoral Researcher with the Department of Electronic Engineering, Fudan University, Shanghai, from 2002 to 2004, where he was a Full Professor, in 2010. Since 2016, he has been an Adjunct Professor with the University of Alberta, Canada. He is currently the Head of the Department of Biomedical Engineering and the Deputy Director of the Institute of Biomedical Engineering and Technology, Fudan University. He has been a Principal Investigator of more than 20 projects. In the last decade, he has contributed more than 200 articles and co-authored five books. He holds 47 patents. His research interests include medical ultrasonic diagnosis systems, medical image and signal processing, ultrasonic inspection, and measurement. He was selected for the First WeiMoan Acoustics Award, in 2013, and the Outstanding Young Scientist Grant of NSF China, in 2015. He is also the Vice-President of the Acoustical Society of China (ASC) and the Medical Engineering Integration Association of China and the Chairperson of the Biomedical Ultrasound Engineering Speciality at ASC. He also served as an IEEE IUS Technical Program Committee (TPC) Member, the Chair of the International Symposium on Ultrasonic Characterization of Bone (ISUCB 2024), and the Co-President/Co-Chair of the 2023 International Congress on Ultrasonics in Beijing (2023 ICU BEIJING).

**EDMOND H. M. LOU** (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 1998. He has held appointments as an Adjunct Professor with the Department of Surgery and Radiobiology and the Department of Diagnostic Imaging, University of Alberta, where he is currently a Full Professor with the Department of Electrical and Computer Engineering and the Department of Biomedical Engineering. His key research is to improve the assessment and treatment effectiveness for children with spinal deformities. He has applied artificial intelligence in X-ray and ultrasound images to assist scoliosis monitoring, and wearable computers to assist orthotic treatment for scoliosis. He is an Associate Fellow of the Scoliosis Research Society.

**LAWRENCE H. LE** received the Ph.D. degree in Earth physics and the M.B.A. degree in finance and technology commercialization from the University of Alberta, Edmonton, AB, Canada, in 1991 and 1999, respectively. He held the Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship at Schlumberger-Doll Research, Ridgefield, CT, USA. He started his medical physics residency training at the Department of Radiology and Diagnostic Imaging (DRDI), University of Alberta, in 1994. He joined DRDI, University of Alberta, as a Clinical Academic Staff, and Capital Health, Edmonton, as a Clinical Medical Physicist, in 2000. He is currently a Clinical Professor of the Graduate Program, DRDI, University of Alberta, and a Senior Medical Physicist with Alberta Health Services, Edmonton. He is also a Senior Visiting Scholar with the Center for Biomedical Engineering, Fudan University, Shanghai, China. His research interests include ultrasound imaging, signal and image processing, wave propagation modeling and inversion, and machine learning. He is a member of American Association of Physicists in Medicine (AAPM) and Canadian Organization of Medical Physicists (COMP).