Received: 3 June 2024 | Revised: 10 November 2024 | Accepted: 14 November 2024

DOI: 10.1002/mp.17562

RESEARCH ARTICLE

MEDICAL PHYSICS

Magnetic resonance image denoising for Rician noise using a novel hybrid transformer-CNN network (HTC-net) and self-supervised pretraining

Shengjie Shi[1] | Cheng Wang[1,2,3] | Sa Xiao[1,2] | Haidong Li[1,2] | Xiuchao Zhao[1,2] | Fumin Guo[4] | Lei Shi[1,2] | Xin Zhou[1,2,5]

[1]Key Laboratory of Magnetic Resonance in Biological Systems, State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, National Center for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences–Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

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

[3]School of Physics and Optoelectronic Engineering, Yangtze University, Jingzhou, China

[4]Wuhan National Laboratory for Optoelectronics, Department of Biomedical Engineering, Huazhong University of Science and Technology, Wuhan, China

[5]Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Haikou, China

Correspondence
Xin Zhou, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, 30 West Xiaohongshan, Wuhan, China.
Email: xinzhou@wipm.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2022YFC2410000; National Natural Science Foundation of China, Grant/Award Numbers: 82127802, 21921004; Key Research Program of Frontier Sciences, Grant/Award Number: ZDBS-Y-JSC004; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDB0540000; Hubei Provincial Key Technology Foundation of China, Grant/Award Numbers: 2021ACA013, 2023BAA021; Natural Science Foundation of Hubei Province, Grant/Award Numbers: 2023AFB1061, 2024AFB284; China Postdoctoral Science Foundation, Grant/Award Number: 2022M713218

Abstract

Background: Magnetic resonance imaging (MRI) is a crucial technique for both scientific research and clinical diagnosis. However, noise generated during MR data acquisition degrades image quality, particularly in hyperpolarized (HP) gas MRI. While deep learning (DL) methods have shown promise for MR image denoising, most of them fail to adequately utilize the long-range information which is important to improve denoising performance. Furthermore, the sample size of paired noisy and noise-free MR images also limits denoising performance.

Purpose: To develop an effective DL method that enhances denoising performance and reduces the requirement of paired MR images by utilizing the long-range information and pretraining.

Methods: In this work, a hybrid Transformer-convolutional neural network (CNN) network (HTC-net) and a self-supervised pretraining strategy are proposed, which effectively enhance the denoising performance. In HTC-net, a CNN branch is exploited to extract the local features. Then a Transformer-CNN branch with two parallel encoders is designed to capture the long-range information. Within this branch, a residual fusion block (RFB) with a residual feature processing module and a feature fusion module is proposed to aggregate features at different resolutions extracted by two parallel encoders. After that, HTC-net exploits the comprehensive features from the CNN branch and the Transformer-CNN branch to accurately predict noise-free MR images through a reconstruction module. To further enhance the performance on limited MRI datasets, a self-supervised pretraining strategy is proposed. This strategy employs self-supervised denoising to equip the HTC-net with denoising capabilities during pretraining, and then the pre-trained parameters are transferred to facilitate subsequent supervised training.

Shengjie Shi and Cheng Wang contributed equally to this study.

Results: Experimental results on the pulmonary HP ^{129}Xe MRI dataset (1059 images) and IXI dataset (5000 images) all demonstrate the proposed method outperforms the state-of-the-art methods, exhibiting superior preservation of edges and structures. Quantitatively, on the pulmonary HP ^{129}Xe MRI dataset, the proposed method outperforms the state-of-the-art methods by 0.254–0.597 dB in PSNR and 0.007–0.013 in SSIM. On the IXI dataset, the proposed method outperforms the state-of-the-art methods by 0.3–0.927 dB in PSNR and 0.003–0.016 in SSIM.

Conclusions: The proposed method can effectively enhance the quality of MR images, which helps improve the diagnosis accuracy in clinical.

KEYWORDS

deep learning, hyperpolarized gas MRI, image denoising, MRI, self-supervised

1 INTRODUCTION

Magnetic resonance imaging (MRI) is a widely used, nonradioactive imaging technique that could provide valuable structural and functional information for both scientific research and clinical diagnosis. Structural MRI facilitates visualization of tissue and organ anatomy with high levels of detail. For example, in brain structural MRI, distinctions between white matter, gray matter, and cerebrospinal fluid enable identification of lesions. Functional MRI, on the other hand, reveals physiological activity of tissues and organs. For instance, hyperpolarized (HP) ^{129}Xe MRI provides lung functional information regarding gas-gas and gas-blood exchange that are highly clinically relevant for lung disease diagnosis.[1,2] However, noise arising from electronics and scanned subjects may decrease the accuracy of quantitative MR measurements and subsequently affect accurate clinical decision making.[3,4] Therefore, denoising is a crucial step in MR imaging workflow.

Many traditional MR image denoising methods have been developed based on filtering or to perform in transformation domain. For example, non-local mean (NLM) filtering eliminates noise by nonlocal averaging of all signal intensities in an image.[5] This approach was adopted by Manjón et al. to denoise MR images with Rician noise, leading to unbiased NLM (UNLM) filtering.[6] Manjón et al. developed a prefiltered rotation invariant NLM (PRINLM) approach by taking advantages of self-similarity properties and sparseness in MR images.[7] Denoising has also been performed in transform domain by transforming spatial images into other domains, and so forth the Fourier, wavelet, and curvelet domain.[8] Dabov et al. combined the concepts of domain transform and nonlocal similarity and developed an image denoising algorithm denominated block-matching and 3-D filtering (BM3D), which was extended to BM4D by Maggioni et al.[9,10] While these methods can provide reasonable denoising performance, they rely on explicit modeling of image and noise patterns, which is difficult to describe and implement for noisy MR images in clinical workflow.[11] Additionally, most of these approaches involve noise estimation, time-consuming optimizations, and manual parameter tuning, resulting in a lack of adaptability.[12]

Recently, deep learning (DL)-based data-driven denoising methods have demonstrated superior performance compared to traditional methods. Zhang et al. proposed a feed-forward denoising convolutional neural network (DnCNN) to eliminate Gaussian noise by utilizing batch normalization and residual learning.[13] Jiang et al. and You et al. improved the DnCNN by tailoring the architecture to effectively handle Rician noise that is common in MR images.[14,15] Ran et al. introduced a residual encoder-decoder Wasserstein generative adversarial network (RED-WGAN) with mean squared error (MSE) and adversarial losses to alleviate the effect of over-smoothing.[16] However, these denoising methods are based on convolutional neural network (CNN), which is known to be limited in capturing long-range information.[11] While capturing long-range information enables the utilization of contextual image information to maintain the continuity of anatomical structures, leading to more reasonable denoising results. To address this issue, Xu et al. proposed a deep adaptive blending network (DABN) that utilized a residual dense block with a large receptive field,[11] and Aetesam and Maji employed dilated convolutions to expand the receptive field of the network.[17] Although these CNN-based somewhat improved the capture range, utilization of such information remains suboptimal because of the inherent limitations of the CNN architectures.[18]

In contrast to CNN, Transformer represents another DL model that is particularly effective at capturing and utilizing long-range information.[19] This capability has led to its widespread adoption in the field of natural language processing. Although Transformer is rarely used in computer vision, some forward-looking studies have demonstrated numerous promise, for example, in natural image denoising.[18,20,21] However, transformer-based methods typically rely on large-scale datasets, restricting their application in the medical domain.[22] Besides, the transformer was reported to have limitations in locality, which could result in the loss of high-frequency details and over-smoothing in denoised images.[21]

Another common challenge in DL-based MR image denoising methods is the requirement of a large number of paired noisy and noise-free images for training, which is difficult to acquire. Self-supervised denoising offers a potential solution by utilizing only noisy images for training,[23–26] but this technique is primarily effective with zero-mean noise and unsuitable for MR images with non-zero mean noise.[27]

To address these challenges, we propose a novel hybrid Transformer-CNN network (HTC-net) with a self-supervised pretraining strategy for MR image denoising. The HTC-net leverages the advantages of Transformer and CNN to improve image denoising performance by capturing long-range information using Transformer while maintaining locality using CNN. In addition, we employ a residual fusion block (RFB) to effectively aggregate feature maps at different resolutions by eliminating the semantic divergence between Transformer and CNN. The pretraining strategy employes self-supervised learning to reduce the need for large-scale paired MR images. Specifically, the strategy performs self-supervised denoising during pretraining and facilitates subsequent supervised training for denoising through transfer learning. Meanwhile, the bias introduced by nonzero mean noise in MR images during pretraining can be corrected by supervised training. To validate the effectiveness of our proposed method, we conduct extensive experiments encompassing both functional and structural MRI datasets. The contributions to this work can be summarized as follows:

1. We propose a novel HTC-net for MR image denoising, which provides the capability of capturing long-range information and maintaining locality.
2. We propose an effective RFB to aggregate hybrid features extracted by Transformer and CNN encoders.
3. We propose a self-supervised pretraining strategy, which further improves algorithm performance and reduces the requirements of paired noisy and noise-free MR images.
4. The proposed method outperforms several state-of-the-art methods for MR image denoising in the lung and brain.

2 | MATERIALS AND METHODS

2.1 | Problem formulation

The main source of noise in MRI is thermal noise, which originates from the scanned object. The variance of thermal noise can be explained as the combined noise variances from independent stochastic processes related to the body, the coil, and the electronics.[4] During MR data acquisition, the raw k-space data are complex values, and both the real and imaginary components are corrupted by Gaussian noise. The complex MR image is then reconstructed from k-space data by inverse Fourier transform. Due to the linearity and orthogonality of the Fourier transform, it preserves the Gaussian characteristics of the noise. Hence, the magnitude MR image can be characterized as follows:

$$x = \sqrt{(y\cos\varphi + n_r)^2 + (y\sin\varphi + n_I)^2} \tag{1}$$

x denotes a noisy MR image, y represents a noise-free MR image, φ indicates the phase, and n_r and n_I are independent and identically distributed Gaussian noise. This is a nonlinear mapping which transforms the Gaussian distribution into a Rician distribution.[28]

MR image denoising is to restore the noise-free MR image y from the noisy MR image x. This process involves finding a mapping f: $x \rightarrow \hat{x}$, where $\hat{x}$ represents the denoised version of x, by minimizing the error between $\hat{x}$ and y:

$$\arg\min_f \|y - \hat{x}\| \tag{2}$$

In this work, our method focuses on magnitude MR images and Rician noise. The mapping f is the network we propose.

2.2 | HTC-net

Figure 1 illustrates the architecture of the proposed HTC-net, which consists of a CNN branch and a Transformer-CNN branch. The CNN branch is exploited to extract the local features. Meanwhile, the Transformer-CNN branch comprises two encoders in parallel to effectively capture long-range information. Let $x \in R^{C,\,H,\,W}$ be an input noisy MR image as input, where C, H, and W denote the number of channels, height, and width of x, respectively. The corresponding denoised MR image $\hat{x} \in R^{C,\,H,\,W}$ can be generated by the proposed network as follows:

$$\hat{x} = f(x) = R(C(x), TC(x)) \tag{3}$$

$R(\cdot)$, $C(\cdot)$, and $TC(\cdot)$ denote the reconstruction module, CNN branch and Transformer-CNN branch respectively.

2.2.1 | CNN branch

In the CNN branch, a series of convolutional layers are employed to extract local features from noisy MR images. Convolution operations are known for their effectiveness in capturing local features, making this branch particularly suitable for enhancing locality within the image. To minimize the potential loss of information caused by downsampling operations, the CNN branch is

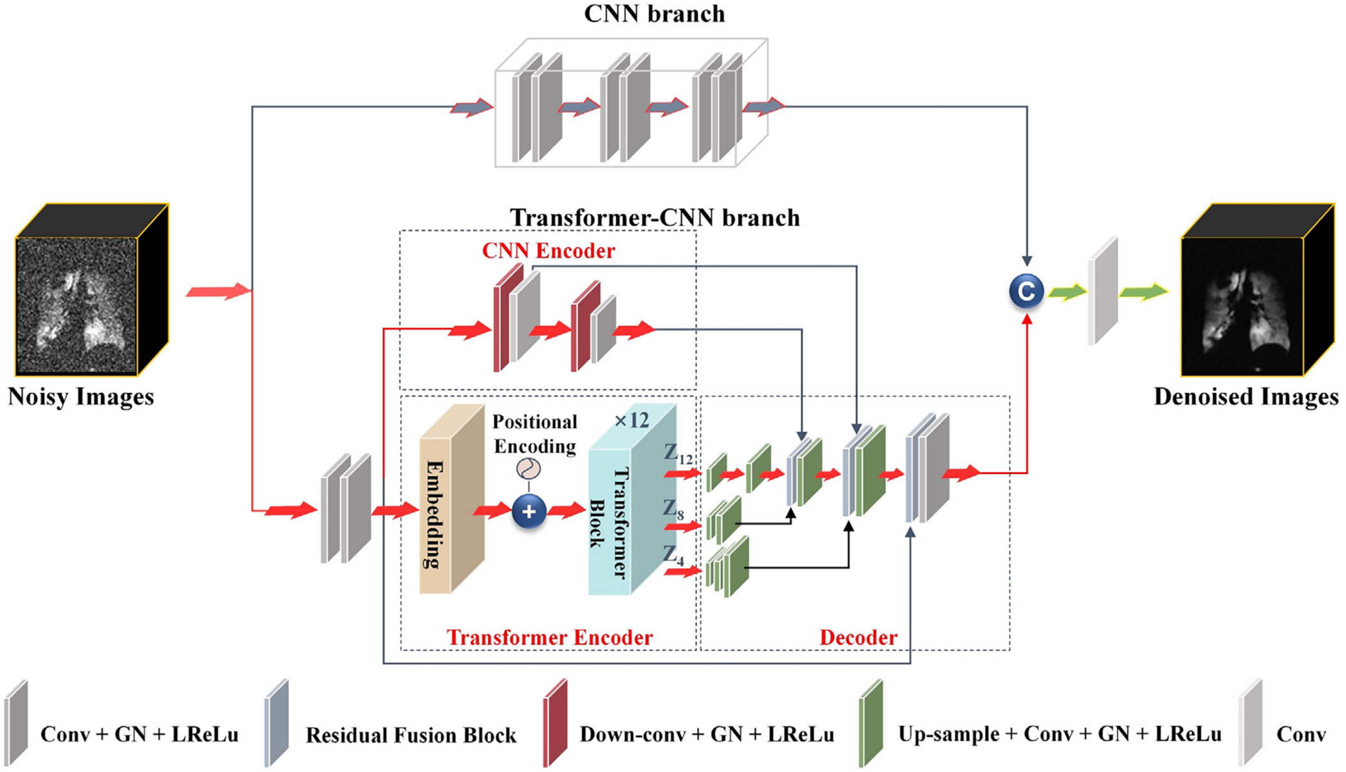

FIGURE 1 The overview of the proposed HTC-net, which consists of the CNN branch and Transformer-CNN branch. The CNN branch comprises a series of convolutional layers, which generate feature maps with the same size as the input image. The Transformer-CNN branch integrates a Transformer encoder, a CNN encoder, and a decoder. The encoders are designed to work in parallel to extract long-range information and local features. Then RFBs are employed to aggregate these hybrid features at different resolution during decoding process. Finally, the hybrid features of both the CNN branch and Transformer-CNN branch are aggregated to obtain the denoised MR image. HTC-net, hybrid transformer-CNN network; RFBs, residual fusion blocks.

designed to keep the sizes of feature maps the same as the input image. This ensures that the spatial information encoded in the feature maps remains intact and aligns with the original input, facilitating accurate and efficient feature extraction. In this work, the CNN branch comprises six layers of 3 × 3 convolutions with stride 1 × 1. The output of the CNN branch is as follows:

$$X_c = C(x) \tag{4}$$

$X_c \in R^{N,\ H,\ W}$, N denotes the number of channels of X_c, which is set to 64 in this work.

2.2.2 | Transformer-CNN branch

In the Transformer-CNN branch, the encoder in Transformer network is employed to enhance the ability of the network in capturing long-range information, while the CNN encoder enhances local feature extraction capabilities. Given a noisy MR image x, the Transformer-CNN branch first uses two 3 × 3 convolutional layers to extract shallow features $F_0 \in R^{N,\ H,\ W}$, which is favorable for early visual processing.[18] The shallow features F_0 are then entered into a Transformer encoder and a CNN encoder to extract deep features.

Transformer encoder To achieve accurate denoising results, it is essential to incorporate the long-range information of the MR images during the denoising process. Here, Transformer encoder is used to extract deep features from shallow features F_0 to enhance denoising performance by leveraging long-range information in the input image. In particular, shallow features F_0 are split into nonoverlapping patches and then mapped into an embedding space. The Transformer encoder is utilized to extract long-range information from these embeddings. Specifically, the embedding layer first downsamples shallow features F_0 by applying stacked convolutional layers:

$$F_0' = D(F_0) \tag{5}$$

D is a stack of convolutional layers, $F_0' \in R^{N',\ H/M,\ W/M}$ are downsampled feature maps with M denoting the downsampling factor. F_0' are split into N_P nonoverlapping patches of size (P, P) with $N_P = HW/(M^2P^2)$, which are further flattened to $N^{'}P^2$-dimensional vectors. After that, the vectors are embedded onto an N_D-dimensional vectors using trainable linear projections[29]:

$$z_0 = \left[F_0^1 P_E; F_0^2 P_E; \dots ; F_0^{N_P} P_E\right] \tag{6}$$

$z_0 \in R^{N_P, N_D}$ are the patch embeddings, P_E is the trainable linear projection, $F_0^p \in R^{N' P^2}$ is the pth patch. These patch embeddings are added with a learnable positional encoding P_E^{pos} and then fed into the Transformer encoder consisting of L Transformer blocks[29]:

$$z_0 = z_0 + P_E^{pos} \tag{7}$$

$$z_l = TB(z_{l-1}) \tag{8}$$

z_l are the output of the lth Transformer block, $TB(\cdot)$ denotes Transformer block. The outputs of the intermediate Transformer blocks are saved and utilized in the subsequent decoding process at different resolutions. By incorporating these intermediate features, we aim to enhance the restoration of details of the MR image. This mechanism allows the decoder to leverage not only the high-level global information but also the finer features captured in the intermediate stages of the encoding process, leading to potentially more comprehensive and accurate denoising results. In this work, M, N', N_D, and L are set to 16, 1024, 768, and 12 respectively. The outputs of the 4th, 8th, and 12th Transformer block are utilized to recover noise-free MR images.

CNN encoder Although the Transformer encoder effectively utilizes long-range information to achieve reasonable denoising results, it has been reported to have a limitation in locality, resulting in potential loss of fine structures such as small lesion regions.[21] To address this issue, a CNN encoder that consists of multiple convolutional layers is employed. This CNN encoder is designed to enhance locality by capturing a hierarchy of localized features. By parallelizing the CNN encoder and the Transformer encoder, we aim to compensate for the locality limitations inherent in the Transformer. This architecture enables the HTC-net to effectively capture both global context and fine-grained local details, thereby improving the overall denoising performance.

In contrast to the CNN branch, the CNN encoder provides a hierarchy of local features, and these multilevel features are fused with features extracted by Transformer encoder. This process involves information loss caused by downsampling operations, which can be compensated by the CNN branch. As shown in Figure 1, the CNN encoder incorporates 2 contraction steps each involves halving the spatial resolution and doubling the channel dimension of the input features. Similar to the Transformer encoder, the output of each contraction step is utilized in subsequent decoding processes to restore MR image details:

$$F_C = CE(F_0) \tag{9}$$

$F_C \in R^{N\times d,\ H/d,\ W/d}$ denote feature maps extracted by CNN encoder, $CE(\cdot)$ denotes CNN encoder. In this work, d is set to 2 and 4 for the two contraction steps, respectively.

Decoder After processing the shallow features F_0, a decoder is utilized to aggregate and decode the global and local features generated by the Transformer encoder and the CNN encoder. To facilitate information flow and aid in the fusion of features, skip connections are utilized between the decoder and encoders. The layers with skip connections in the encoder include three steps:

1. upsampling the embeddings of Transformer encoder. To effectively integrate the features of the two encoders, the Transformer embeddings are first reshaped into 2D feature maps. Subsequently, the feature maps are upscaled to align with the dimensions of their corresponding CNN feature maps. Specifically, the Transformer embeddings z_4, z_8, and z_{12} are processed using 3, 2, and 1 upsampling operations, respectively. This process ensures alignment and fusion of feature maps generated by both the Transformer and CNN encoders.
2. aggregating the hybrid features. In this step, a RFB is employed to aggregate the upsampled Transformer feature maps, the CNN feature maps, and the upsampled feature maps. As shown in Figure 2, the RFB consists of a residual feature processing module and a feature fusion module. In the residual feature processing module, residual blocks are utilized to process the upsampled Transformer feature maps and the CNN feature maps, which can be stated as:

$$F'_T = F_T + RB(F_T, F_C, F_U) \tag{10}$$

$$F'_C = F_C + RB(F_T, F_C, F_U) \tag{11}$$

$F_T \in R^{N\times d,\ H/d,\ W/d}$ denote the upsampled Transformer feature maps, $F_U \in R^{N\times d,\ H/d,\ W/d}$ denote the upsampled feature maps, and $RB(\cdot)$ denotes residual block. Here, the residual blocks are designed to adjust biases for the upsampled Transformer and the CNN feature maps to minimize semantic divergence and facilitate effective feature fusion. In the feature fusion module, hybrid F'_T, F'_C, and F_U are fused by 1 × 1 convolution:

$$F_{fuse} = CN\left(F'_T, F'_C, F_U\right) \tag{12}$$

$CN(\cdot)$ is the 1×1 convolution operation, $F_{fuse} \in R^{N\times d,\ H/d,\ W/d}$ are the fused feature maps.
3. upsampling the fused features. In this step, the fused feature maps are upsampled using bilinear interpolation and convolutional layers with a kernel size 3 × 3 to obtain the next level of upsampled feature maps

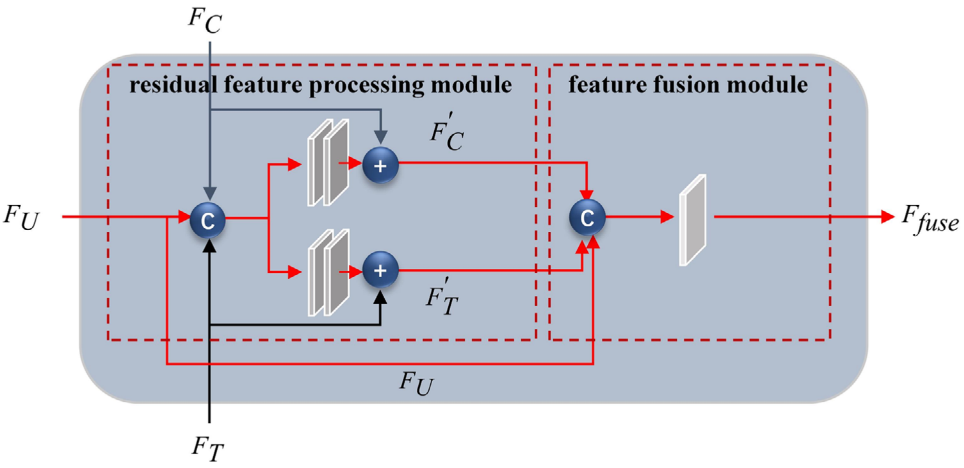

FIGURE 2 The architecture of the proposed RFB. The RFB consists of a residual feature processing module and a feature fusion module, which is utilized to aggregate the upsampled Transformer feature maps (F_T), the CNN feature maps (F_C) and the upsampled feature maps (F_U). In residual feature processing module, the residual blocks are utilized to process F_T, F_C, and F_U to obtain F'_T and F'_C. Then the fused feature maps (F_{fuse}) are obtained by using a convolution layer to aggregate F'_T, F'_C, and F_U in feature fusion module. CNN, convolutional neural network; RFB, residual fusion block.

$F_U \in R^{N\times d/2,\, 2\times H/d,\, 2\times W/d}$:

$$F_U = U(F_{fuse}) \tag{13}$$

$U(\cdot)$ denotes the upsampling operation.

In the final layer of the decoder, the upsampled feature maps F_U are fused with the shallow feature maps F_0 to generate feature maps $X_{tc} \in R^{N,\, H,\, W}$, the output features of the Transformer-CNN branch.

Reconstruction module The hybrid feature maps of both the CNN branch and Transformer-CNN branch are aggregated to obtain the denoised MR image $\hat{x}$ through a reconstruction module implemented using a 1 × 1 convolution layer.

$$\hat{x} = R(X_c, X_{tc}) \tag{14}$$

2.3 | Pretraining strategy

The availability of paired noisy and noise-free MR images is limited, which restricts the denoising performance of DL methods. Here, to further improve the denoising performance of the proposed network, we introduce a pretraining strategy based on the self-supervised denoising. The self-supervised loss can generally be expressed as:

$$\min_{\theta} \boldsymbol{E}_x \|f_\theta(x) - x\|^2 \tag{15}$$

x denotes the noisy image, f_θ denotes the denoising network. Based on the idea of J-invariance,[24] when f_θ is J-invariant, the self-supervised loss function can be written as:

$$\boldsymbol{E}_x \|f_\theta(x) - x\|^2 = \boldsymbol{E}_{x,y} \|f_\theta(x) - \bar{x}\|^2 + \boldsymbol{E}_{x,y} \|x - \bar{x}\|^2 \tag{16}$$

$\bar{x} = \mathrm{E}_{x|y}(x)$, y is the noise-free image. If the noise is zero mean, minimizing a self-supervised loss across J-invariant functions is equivalent to minimizing a supervised loss.

However, it is important to note that the Rician noise present in MR images is not zero mean. This means that noise-free images cannot be directly recovered from noisy MR images using the self-supervised denoising method. The mean of the MR image is

$$\bar{x} = \int_0^\infty x \frac{x}{\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) I_0\left(\frac{yx}{\sigma^2}\right) dx \tag{17}$$

σ denotes standard deviation of Gaussian distribution and I_0 denotes the zeroth order Modified Bessel function. Although $\bar{x}$ is different from noise-free image y, the denoised image $\bar{x}$ has no random fluctuations and can be uniquely determined by y and σ.[30] In other words, $\bar{x}$ can be regarded as a biased denoising result that provides some structural information of the noise-free MR images.

Consequently, to further exploit the information of noisy images and improve the network denoising performance, we propose a pretraining strategy based on the self-supervised denoising. This strategy aims to provide suitable initial parameters and a certain level of denoising ability for supervised training. As shown in Figure 3, the proposed pretraining strategy includes three steps:

1. **randomly masking pixels in noisy MR images;** we randomly mask some pixels in noisy MR images.

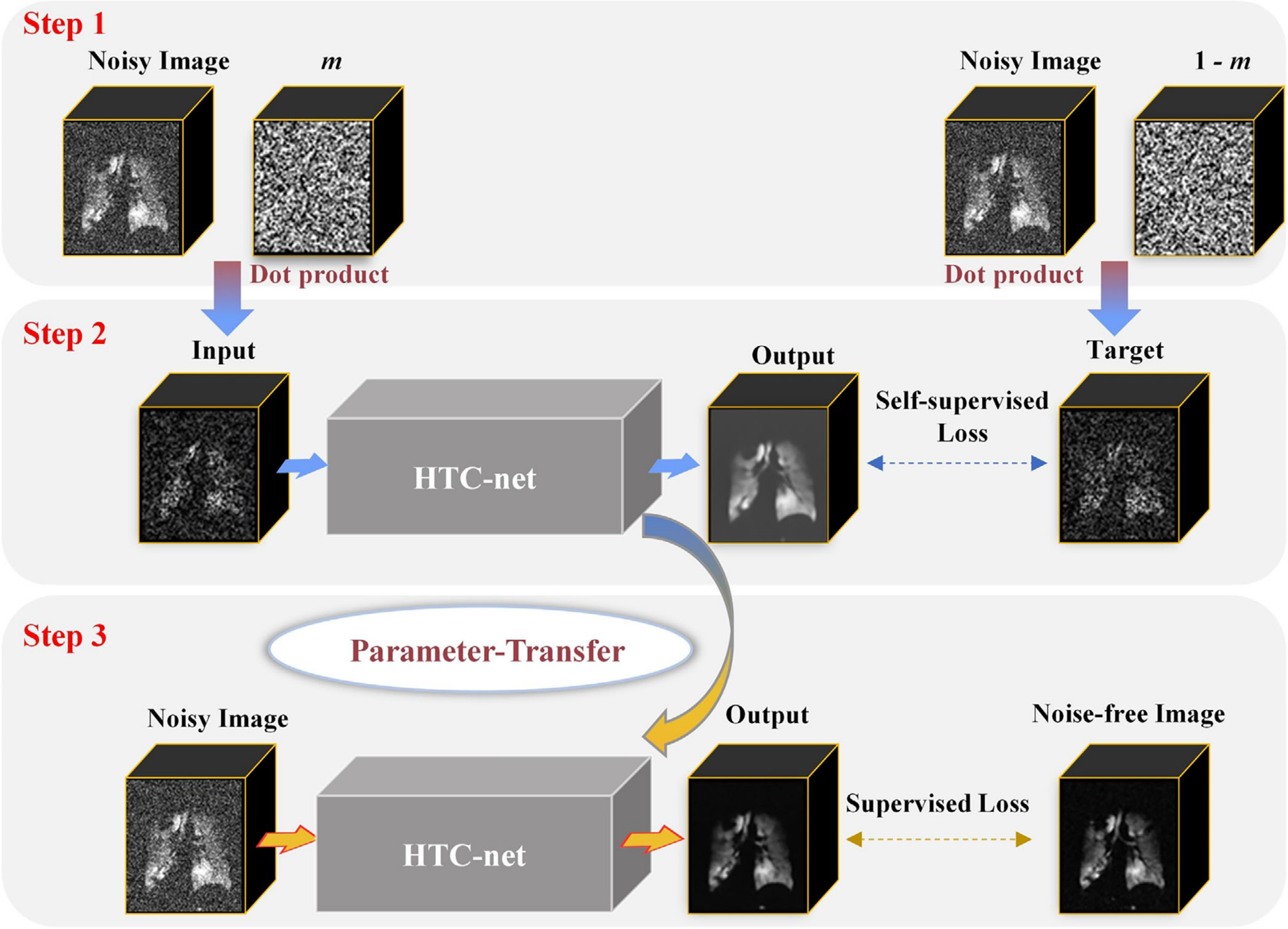

FIGURE 3 The workflow of the pretraining strategy based on self-supervised denoising. The training strategy includes three steps: (1) randomly masking pixels of noisy MR images. This step involves using a randomly generated mask *m* to create paired noisy MR images, forming the masked dataset; (2) pretraining HTC-net with the masked dataset. The input of HTC-net is the image masked by *m* and the target is the image masked by 1 - *m*; (3) transferring the parameters in supervised learning. Transferring the pre-trained parameters to supervised training with paired noisy and noise-free MR images. HTC-net, hybrid transformer-CNN network; MR, magnetic resonance.

The process of random masking follows a Bernoulli distribution with probability *p*. Then the masked and unmasked pixels form paired with noisy MR images in the masked dataset. In this work, *p* is set to 0.5.

2. **pretraining HTC-net with the masked dataset;** pretraining is performed with the constructed masked dataset. Here, the HTC-net predicts the masked pixels by utilizing the unmasked pixels in the noisy MR image, which makes the network *J*-invariant. Subsequently, the pretrained network gains the ability of extracting image features and denoising. This pretraining process provides appropriate initial parameters for the subsequent supervised training for denoising.
3. **transferring the network parameters in supervised learning;** the parameters obtained from pretraining procedure are utilized to initialize the supervised training that aims to refine the denoising capability using paired noisy and noise-free MR images.

2.4 | Loss function

The loss function consists of self-supervised loss for pretraining and supervised loss in supervised training.

The self-supervised loss function can be stated as:

$$\min_{\theta} \|(1-m)[f_{\theta}(m \cdot x) - (1-m) \cdot x]\|_2^2 \tag{18}$$

m is a generated mask.

The supervised loss function is calculated as the absolute differences between denoised image $f_{\theta}(x)$ and ground truth noise-free image *y*:

$$\min_{\theta} \|f_{\theta}(x) - y\|_1 \tag{19}$$

2.5 | Datasets

To validate the performance of our method, we performed experiments on functional and structural MRI datasets.

Functional MRI dataset: pulmonary HP ^{129}Xe MRI allows visualization of lung ventilation function, which is valuable for clinical diagnosis of pulmonary diseases. In this study, pulmonary HP ^{129}Xe ventilation images were obtained from 85 subjects. All experiments were conducted on a 1.5 T whole-body MRI scanner (Avanto, Siemens Medical Solutions) and were approved by the local Institutional Review Board. Based on the Rb-^{129}Xe spin-exchange optical pumping (Rb-^{129}Xe SEOP) method, enriched ^{129}Xe gas was polarized using a commercial xenon polarizer (verImagin Healthcare, Wuhan). After inhaling the gas mixture from functional residual capacity, subjects held their breath for data acquisition.

The MRI parameters for HP ^{129}Xe imaging were: matrix size = 96 × 84, repetition time/echo time (TR/TE) = 4.2/1.9 ms, field of view = 384 × 336 mm^2, slice thickness = 8 mm, bandwidth = 38.4 kHz, number of slices = 24, 3D bSSFP sequence, flip angle (FA) = 10°, scan time = 8.4 s.

We selected images with signal-to-noise ratios (SNRs) greater than 15 from the acquired HP ^{129}Xe MRI data, and finally obtained 1059 HP ^{129}Xe images for experiments.[31] These images were padded to 96 × 96. From this dataset, we randomly chose 847 images for training and reserved 212 images for testing. Training images were augmented with horizontal flips and the 90°, 180°, and 270° rotations. When the SNRs are less than 8, HP ^{129}Xe images require reacquisition to ensure sufficient image quality.[32] Therefore, Rician noise was added to these images to obtain noisy HP ^{129}Xe images with SNRs ranging from 8 to 15 and from 5 to 8.

Structural MRI dataset: the structural MR images used in this work were from the public IXI dataset (http://brain-development.org/ixidataset/). The IXI dataset includes T1, T2, and proton density (PD)-weighted images from three different hospitals in London. Hammersmith dataset is a subset of IXI dataset, and the data was acquired using a Philips 3T system. We chose the T1-weighted images among them and the T1-weighted MRI parameters in Hammersmith dataset were TR/TE = 9.6/4.6 ms, number of phase encoding steps = 208, reconstruction diameter = 240.0, FA = 8°.

We removed the anterior and posterior slices from each MRI volume from the IXI dataset, resulting in 5000 2D images. Then we randomly selected 4000 images from them as the training set and 1000 images as the test set. The size of each image is 256 × 256. To maintain consistency with HP ^{129}Xe MRI, we add Rician noise into these images to obtain noisy MR images with SNRs ranging from 8 to 15 and from 5 to 8.

Under-sampled MRI dataset: two-fold under-sampled pulmonary HP ^{129}Xe MR images were acquired as the real noisy MR images by using a variable density Cartesian sampling pattern in a phase-encoding direction. The matrix size of each image is 96 × 84, and a total of 620 under-sampled images from 85 subjects were obtained for testing. These images were also padded to 96 × 96 in testing process.

2.6 | Implementation details

Our network was implemented using PyTorch framework. Training and testing were performed on a workstation with an NVIDIA GTX 3090 of 24 GB memory and an Intel Xeon(R) W-2245 CPU. AdamW optimizer with β_1 of 0.9 and β_2 of 0.999 was adopted.[33] In the self-supervised pretraining and supervised training phase, the learning rates were set to 5e-4 and 1e-4, respectively, and decayed by half every 50 epochs. The batch size was set to eight and the training was stopped after 60 epochs. To demonstrate the advantages of the proposed network architecture, we also conducted supervised training on the network without pretraining, and the training was stopped after 100 epochs.

2.7 | Performance evaluation

To validate the effectiveness of the HTC-net with pretraining (called P-HTC-net), it was compared with traditional methods [NLM,[5] UNLM[6]] and DL methods [DnCNN,[13] wider denoising neural network (WDNN),[15] image restoration model based on Swin Transformer (SwinIR),[18] HTC-net]. Quantitative denoising results were evaluated based on peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [34,35]:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{I_{max}^2}{\mathrm{MSE}}\right) \tag{20}$$

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{21}$$

x and y represent the denoised MR image and the noise-free image, respectively, MSE denotes the MSE between x and y, μ_x and μ_y are the mean of x and y, σ_x and σ_y are the standard deviations of x and y, σ_{xy} is the covariance between x and y, c_1 and c_2 are constants.

Ventilation defect percentage (VDP) is an important quantitative metric for assessing pulmonary function in lung MRI, which is defined as the ratio of the ventilation defect volume to the thoracic volume.[1] For pulmonary HP ^{129}Xe ventilation images, accurate segmentation of ventilation defect regions (lung lesions) is essential for calculating VDP.[36] Thus, we undertook a comparison of the segmentation of ventilation defect regions in denoised lung images to further evaluate the accuracy of the denoising results. White matter atrophy serves as a clinical marker for a diverse range of neurological disorders. Accurate segmentation of white matter is necessary to estimate the white matter

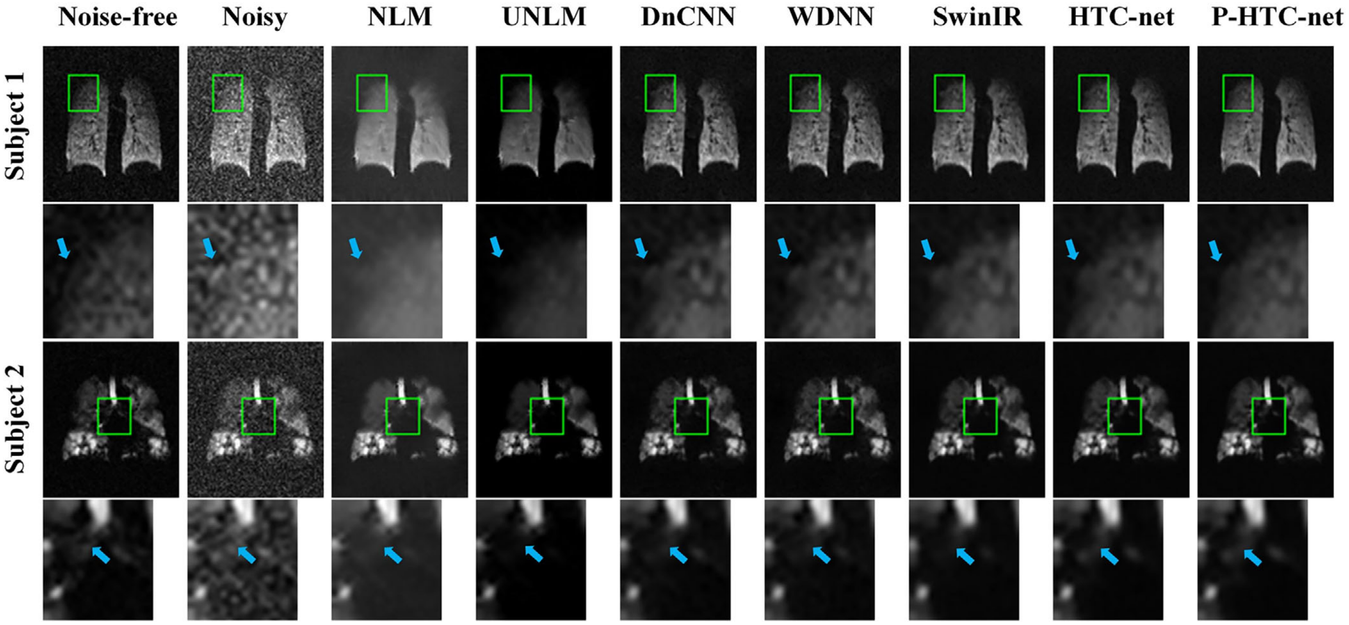

FIGURE 4 The denoising results of two representative subjects on the pulmonary HP ^{129}Xe MRI dataset (SNRs ≥ 8): subject 1 (top two rows) and subject 2 (bottom two rows). The second and fourth rows show the zoomed-in details.

volume.[37] Therefore, we compared the segmentation of white and gray matter in denoised brain images from the IXI dataset to further evaluate the accuracy of the denoising results. Dice score was used to evaluate the quantitative results of segmentation:

$$\mathrm{Dice}\,(S, G) = \frac{2\,|S \cap G|}{|S| + |G|} \tag{22}$$

S represents the segmentation result and G denotes the ground truth.

For under-sampled experiments, we first evaluated the denoising performance using SNR. Then, we evaluated performance through experts scoring due to the lack of ground truth (clean MR images). Two clinical experts with more than 5 years of experience in lung MRI evaluate the quality of these denoised images based on the overall image quality, sharpness, and subjective SNR, with scores ranging from 1 to 5.[38] A higher score means the better image quality.

3 | RESULTS

3.1 | Results on the pulmonary HP ^{129}Xe MRI dataset

Figures 4 and 5 show the representative denoising results of different denoising methods at varying SNR levels. As shown in these figures, traditional denoising methods often result in over-smoothed images, lacking in fine details. Conversely, the DL methods perform better in preserving details. Among them, our method outperforms others in recovering low signal areas of HP ^{129}Xe images while more efficiently reducing background noise. At low SNR levels, our method is still able to provide clear airways and enhanced edge sharpness. In summary, the proposed method outperforms others by yielding visually superior results with more details.

Table 1 presents quantitative results comparing various denoising methods on pulmonary HP ^{129}Xe MRI dataset at different SNR levels. It is evident that all denoising methods contribute to enhancing image quality, with our method demonstrating the highest PSNR and SSIM values at each SNR level (PSNR of 30.022 and 27.888 dB, SSIM of 0.869 and 0.822). Notably, DL methods consistently outperform traditional methods in denoising metrics. Among the compared methods, SwinIR exhibits suboptimal denoising performance, followed by WDNN and DnCNN. In comparison with SwinIR, our HTC-net, even without pretraining, achieves an increase of 0.239 and 0.22 dB in PSNR values within SNR ranges of 8 to 15 and 5 to 8, respectively. With pretraining, the PSNR values of our network further improve by 0.015 and 0.084 dB, respectively. A paired t-test is used to statistically compare the differences between our method and other methods. The corresponding p-values are presented in Table S1, indicating that our method differs significantly from others ($p < 0.05$).

In Figure 6, segmentation results for ventilation defect regions (lesions) in denoised images are illustrated. Notably, images denoised with NLM and UNLM exhibit challenges in accurately segmenting small ventilation defect regions, indicating limitations in detail recovery for HP ^{129}Xe image. In contrast, the segmentation results based on our method are closest to the ground truth

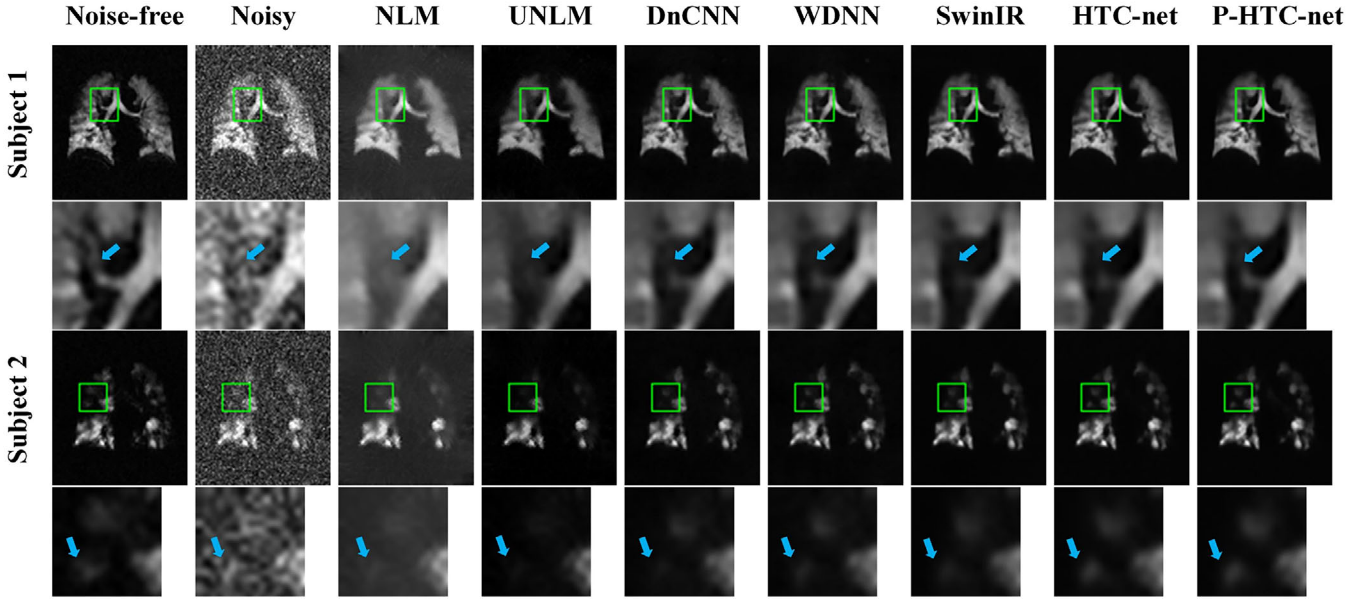

FIGURE 5 The denoising results of two representative subjects on the pulmonary HP ^{129}Xe MRI dataset (SNRs < 8): subject 1 (top two rows) and subject 2 (bottom two rows). The second and fourth rows show the zoomed-in details. MRI, magnetic resonance imaging; SNRs, signal-to-noise ratios.

TABLE 1 Quantitative results of different denoising methods on the pulmonary HP ^{129}Xe MRI dataset in terms of PSNR and SSIM (mean ± standard deviation).

	SNRs = (8, 15)		SNRs = (5, 8)	
Method	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Noisy	17.233 ± 3.120**	0.481 ± 0.128**	10.884 ± 1.703**	0.225 ± 0.062**
NLM	20.241 ± 3.003**	0.498 ± 0.085**	14.352 ± 1.182**	0.324 ± 0.053**
UNLM	23.491 ± 2.898**	0.390 ± 0.140**	22.285 ± 2.476**	0.393 ± 0.153**
DnCNN	29.526 ± 2.284**	0.859 ± 0.053**	27.480 ± 1.836**	0.810 ± 0.061**
WDNN	29.425 ± 2.158**	0.856 ± 0.051**	27.525 ± 1.846**	0.809 ± 0.063**
SwinIR	29.768 ± 2.348**	0.862 ± 0.055**	27.584 ± 1.885**	0.812 ± 0.064**
HTC-net	30.007 ± 2.429*	0.869 ± 0.057	27.804 ± 1.925**	0.820 ± 0.065**
P-HTC-net	**30.022** ± **2.428**	**0.869** ± **0.057**	**27.888** ± **1.925**	**0.822** ± **0.064**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.

Abbreviations: DnCNN, denoising convolutional neural network; HTC-net, hybrid transformer-CNN network; NLM, nonlocal mean; PSNR, peak signal-to-noise ratio; SNRs, signal-to-noise ratios; SSIM, structural similarity index measure; UNLM, unbiased NLM.

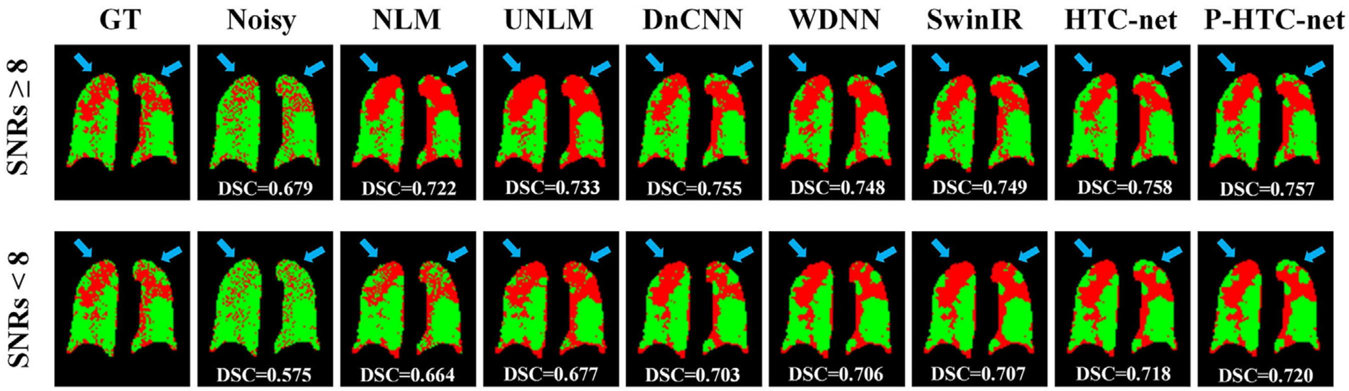

FIGURE 6 The segmentation results on the pulmonary HP ^{129}Xe MRI dataset. The green color denotes the healthy region, and the red color denotes the lesions. The top row is segmentation results of denoised pulmonary HP ^{129}Xe MR images with SNRs ≥ 8, and the bottom row is segmentation results of denoised pulmonary HP ^{129}Xe MR images with SNRs < 8. HP, hyperpolarized; MR, magnetic resonance; MRI, magnetic resonance imaging; SNRs, signal-to-noise ratios.

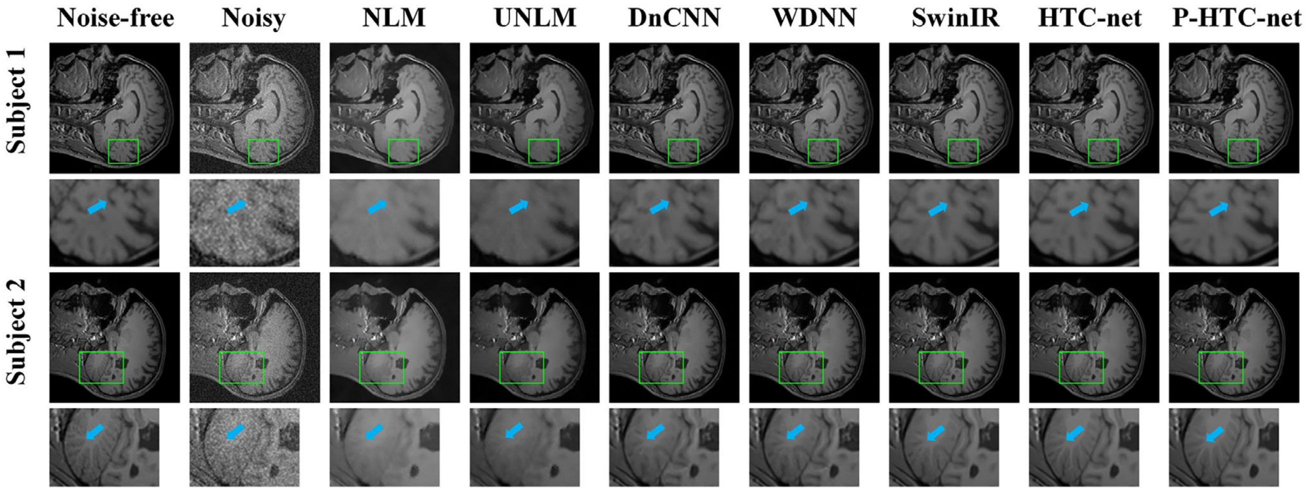

FIGURE 7 The denoising results from two representative subjects on the IXI dataset (SNRs ≥ 8): subject 1 (top two rows) and subject 2 (bottom two rows). The second and fourth rows show the zoomed-in details. SNRs, signal-to-noise ratios.

TABLE 2 Dice scores obtained by various denoising methods on the pulmonary HP ^{129}Xe MRI dataset (mean ± standard deviation).

Method	SNRs = (8, 15)	SNRs = (5, 8)
Noisy	0.680 ± 0.057**	0.567 ± 0.058**
NLM	0.748 ± 0.070**	0.671 ± 0.080**
UNLM	0.744 ± 0.069**	0.687 ± 0.071**
DnCNN	0.769 ± 0.079**	0.710 ± 0.081**
WDNN	0.761 ± 0.086**	0.718 ± 0.073**
SwinIR	0.769 ± 0.078**	0.721 ± 0.074**
HTC-net	0.775 ± 0.080*	0.734 ± 0.072*
P-HTC-net	**0.784** ± **0.079**	**0.739** ± **0.075**

Note: The best results are shown in bold. * and ** denote p-value < 0.05 and p-value < 0.001, respectively.

Abbreviations: DnCNN, denoising convolutional neural network; HTC-net, hybrid transformer-CNN network; NLM, nonlocal mean; SNRs, signal-to-noise ratios; UNLM, unbiased NLM.

and effectively recovers more small ventilation defect regions compared to other methods. Table 2 lists the Dice scores obtained by various denoising methods on the pulmonary HP ^{129}Xe MRI dataset. The proposed method obtains the highest Dice scores (0.784 and 0.739), which outperforms other DL methods by 2%–4%. The corresponding p-values are presented in Table S1, indicating that our method differs significantly from others ($p < 0.05$). We also compare the correlation between the VDP obtained by various methods and the VDP of noise-free images, as shown in Figures S1 and S2.

3.2 | Results on the IXI dataset

Figures 7 and 8 present the representative results of denoising methods on the IXI dataset with SNRs ranging from 8 to 15 and 5 to 8. It can be observed that our method not only recovers sulcus and gyrus in brain MR images, but also provides a clear boundary between gray matter and white matter. Moreover, our method provides clearer details in the zoomed-in images of the cerebellum. It is evident that our method is able to restore more fine information, preserving clear shapes, and edges. Comparatively, other denoising methods tend to lose details or introduce false features.

Table 3 presents the quantitative results of various denoising methods on the IXI dataset. DL methods obtain better denoising performance than traditional methods. In comparison to other methods, our method obtains the highest quantitative metrics (PSNR of 35.507 and 32.275 dB, SSIM of 0.973 and 0.948). Specifically, in the dataset with SNRs ranging from 8 to 15 and 5 to 8, the PSNR values that our method achieves are 0.3 and 0.587 dB higher than SwinIR (the method with the highest PSNR among others) respectively. Even without self-supervised pretraining, the proposed network still outperforms other methods in terms of PSNR and SSIM, verifying the superiority of its structure. Table S1 shows that our method differs significantly from others ($p < 0.05$).

Figure 9 displays the segmentation results of white and gray matter in denoised brain images from the IXI dataset. Notably, the segmentation results of noisy images fail to provide clear boundaries between white and gray matter, yielding the worst segmentation results. However, all the denoising methods improve the segmentation results. Among them, the segmentation results of the proposed denoising method are closest to ground truth. Table 4 presents the Dice scores obtained by various denoising methods on the IXI dataset. The proposed method obtains the highest Dice scores (0.895 and 0.853), outperforming other DL methods by 1%–6%. Statistical analysis indicates that our method differs significantly from other methods ($p < 0.05$).

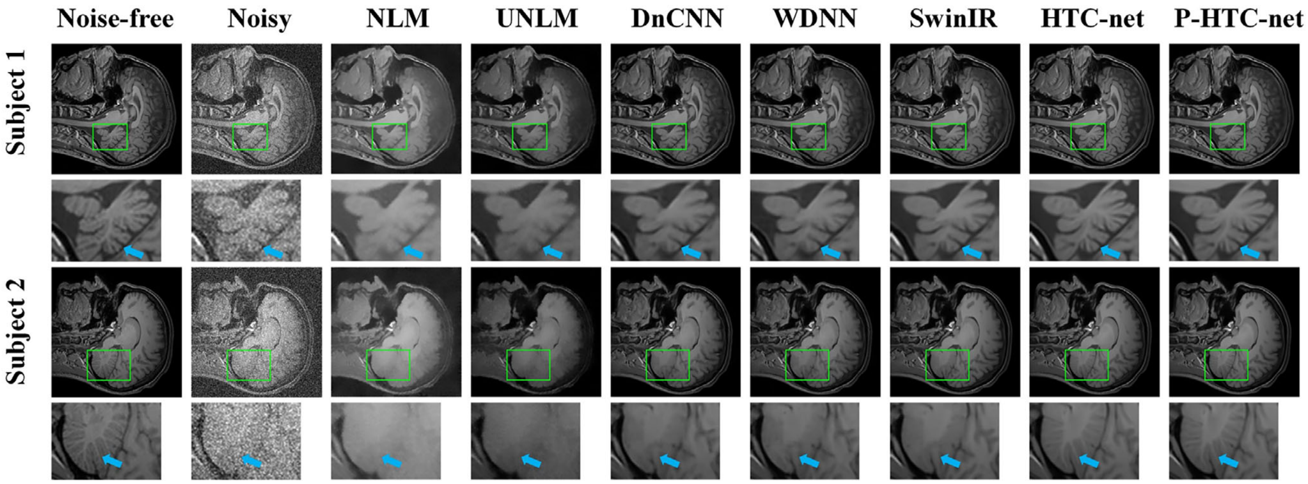

FIGURE 8 The denoising results from two representative subjects on the IXI dataset (SNRs < 8): subject 1 (top two rows) and subject 2 (bottom two rows). The second and fourth rows show the zoomed-in details. SNRs, signal-to-noise ratios.

TABLE 3 Quantitative results of different denoising methods on the IXI dataset in terms of PSNR and SSIM (mean ± standard deviation).

	SNRs = (8, 15)		SNRs = (5, 8)	
Method	PSNR	SSIM	PSNR	SSIM
Noisy	18.585 ± 1.099**	0.465 ± 0.056**	12.278 ± 0.997**	0.244 ± 0.047**
NLM	21.358 ± 1.133**	0.510 ± 0.047**	15.574 ± 1.050**	0.369 ± 0.043**
UNLM	32.065 ± 1.172**	0.930 ± 0.022**	27.855 ± 1.151**	0.832 ± 0.048**
DnCNN	34.813 ± 1.231**	0.967 ± 0.007**	31.348 ± 1.182**	0.932 ± 0.012**
WDNN	34.961 ± 1.233**	0.968 ± 0.007**	31.409 ± 1.162**	0.933 ± 0.012**
SwinIR	35.207 ± 1.245**	0.970 ± 0.006**	31.688 ± 1.210**	0.939 ± 0.011**
HTC-net	35.392 ± 1.238**	0.972 ± 0.006**	32.105 ± 1.203**	0.947 ± 0.010**
P-HTC-net	**35.507 ± 1.243**	**0.973 ± 0.005**	**32.275 ± 1.222**	**0.948 ± 0.010**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.

Abbreviations: DnCNN, denoising convolutional neural network; HTC-net, hybrid transformer-CNN network; NLM, nonlocal mean; PSNR, peak signal-to-noise ratio; SNRs, signal-to-noise ratios; SSIM, structural similarity index measure; UNLM, unbiased NLM.

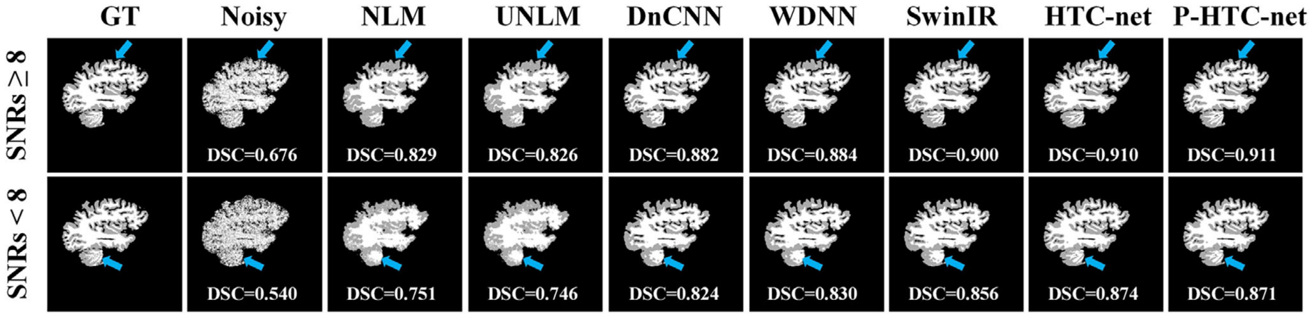

FIGURE 9 The segmentation results on the IXI dataset. The white color denotes the white matter, and the gray color denotes the gray matter. The top row is segmentation results of denoised brain images with SNRs ≥ 8, and the bottom row is segmentation results of denoised brain MR images with SNRs < 8. MR, magnetic resonance; SNRs, signal-to-noise ratios.

3.3 | Results on the real noisy MR images

We present denoised results of various methods on 620 real noisy pulmonary HP ^{129}Xe MR images that are two-fold under-sampled, evaluated using SNR and experts scoring. The DL methods are trained on pulmonary HP ^{129}Xe MRI dataset with synthetic noise. The results are shown in Figures 10 and S3, and Table 5. Figure 10 shows that our method preserves more

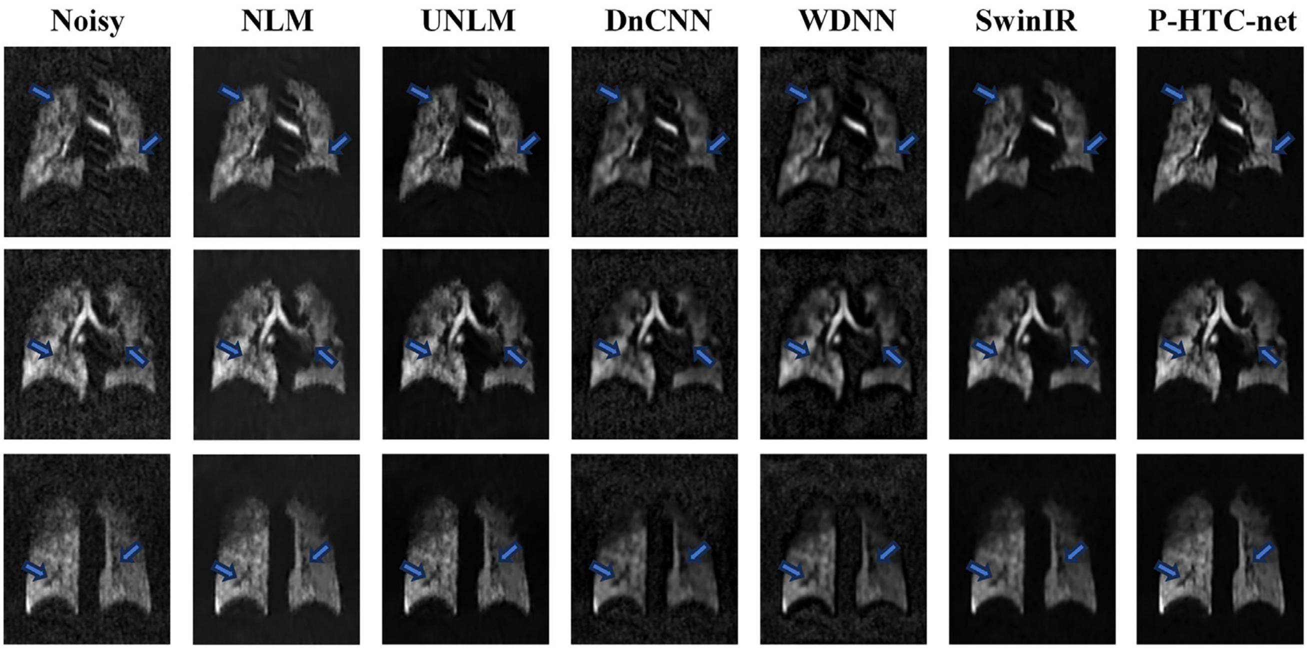

FIGURE 10 The denoising results on the pulmonary under-sampled MRI data. MRI, magnetic resonance imaging.

TABLE 4 Dice scores obtained by various denoising methods on the IXI dataset (mean ± standard deviation).

Method	SNRs = (8, 15)	SNRs = (5, 8)
Noisy	0.643 ± 0.060**	0.515 ± 0.034**
NLM	0.805 ± 0.043**	0.711 ± 0.040**
UNLM	0.808 ± 0.042**	0.714 ± 0.037**
DnCNN	0.872 ± 0.025**	0.805 ± 0.028**
WDNN	0.878 ± 0.025**	0.807 ± 0.030**
SwinIR	0.886 ± 0.024**	0.824 ± 0.029**
HTC-net	0.894 ± 0.023**	0.848 ± 0.029**
P-HTC-net	**0.895 ± 0.023**	**0.853 ± 0.026**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.

Abbreviations: DnCNN, denoising convolutional neural network; HTC-net, hybrid transformer-CNN network; NLM, nonlocal mean; SNRs, signal-to-noise ratios; UNLM, unbiased NLM.

image details, providing clearer ventilation defects, and sharper denoised results. In quantitative comparisons, our method outperforms the state-of-the-art method by 19.4% in SNR (104.89 vs. 87.86), demonstrating better generalization ability. Additionally, as demonstrated in expert evaluations, Figure S3 shows that our method achieves the highest scores in terms of overall image quality, sharpness, and SNR.

3.4 | Ablation experiments

Ablation experiments are conducted on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset. We evaluate various network configurations to evaluate their performance. These include: the CNN branch alone (u), the Transformer-CNN branch without the Transformer encoder (d), and several combinations of them. The combinations are the CNN and Transformer-CNN branches without the Transformer encoder (u + d), adding the Transformer encoder to this mix (u + d + t), further incorporating the RFB (u + d + t + f), and the complete method which includes self-supervised pretraining (u + d + t + f + s). Figure 11 and Figure S5 present subjective results of various network configurations on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset. Among these, the (u + d + t + f + s) yields the clearest details, closely matching the noise-free images. The quantitative results are summarized in Tables 6 and 7. The (u + d + t + f + s) achieves the highest metrics (for the HP ^{129}Xe MRI dataset: PSNR of 30.022 dB and 27.888 dB, SSIM of 0.869 and 0.822; for the IXI dataset: PSNR of 35.507 and 32.275 dB, SSIM of 0.973 and 0.948). Table S2 presents the *p*-values for various network variants compared to our method, demonstrating statistically significant results.

To further evaluate the effectiveness of pretraining strategy, different numbers of noisy MR images are used during the pretraining stage. For the IXI dataset, we augment the self-supervised pretraining by incorporating an additional 4000 noisy MR images. These images are obtained by adding Rician noise to the T1-weighted images from the IXI dataset. For the pulmonary HP ^{129}Xe MRI dataset, we selected additional 753 HP ^{129}Xe images with SNRs less than 15 to augment the self-supervised pretraining. Table 8 presents the quantitative results on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset, respectively. As the number of noisy MR images increases, the denoising metrics obtained

TABLE 5 Quantitative results of various denoising methods on the pulmonary under-sampled MRI data (mean ± standard deviation).

	Noisy	NLM	UNLM	DnCNN	WDNN	SwinIR	P-HTC-net
SNR	18.50 ± 3.04	70.42 ± 17.05	64.83 ± 16.27	17.83 ± 6.02	19.10 ± 8.29	87.86 ± 17.47	**104.89 ± 39.78**
Score	1.43 ± 0.15	1.78 ± 0.38	2.57 ± 0.50	2.35 ± 0.23	2.23 ± 0.13	3.45 ± 0.18	**3.95 ± 0.13**

Note: The best results are shown in bold. Score is obtained by averaging the overall image quality, sharpness, and SNR scores given by experts.
Abbreviations: DnCNN, denoising convolutional neural network; NLM, nonlocal mean; SNRs, signal-to-noise ratios; UNLM, unbiased NLM.

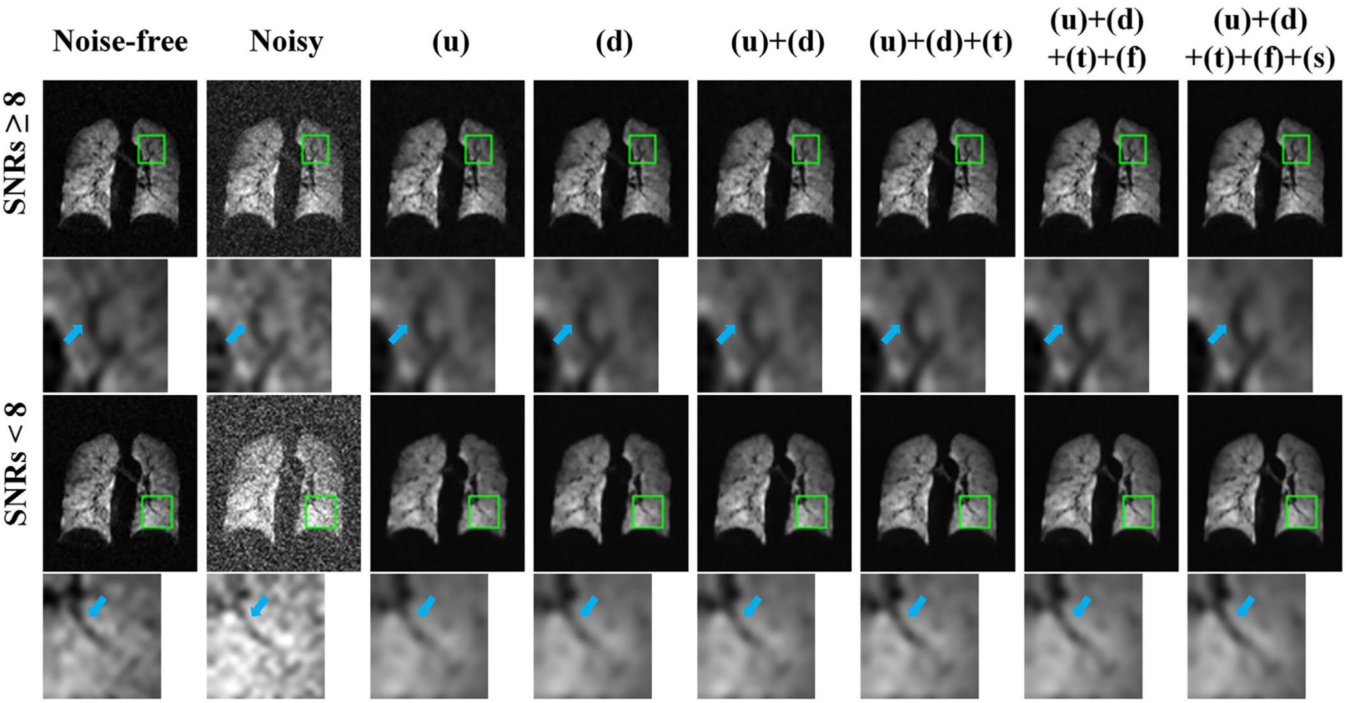

FIGURE 11 The denoising results of various network configurations on the pulmonary HP ^{129}Xe MRI dataset: (u) denoising with only the CNN branch, (d) denoising with only the Transformer-CNN branch without transformer encoder, (t) the transformer encoder, (f) the residual fusion block, (s) the self-supervised pretraining. HP, hyperpolarized; CNN, convolutional neural network; MRI, magnetic resonance imaging.

TABLE 6 Ablation studies on various network configurations on the pulmonary HP ^{129}Xe MRI dataset (mean ± standard deviation): (u) denoising with only the CNN branch, (d) denoising with only the Transformer-CNN branch without Transformer encoder, (t) the Transformer encoder, (f) the RFB, (s) the self-supervised pretraining.

	SNRs = (8, 15)		SNRs = (5, 8)	
Method	PSNR	SSIM	PSNR	SSIM
(u)	29.513 ± 2.285**	0.856 ± 0.053**	27.527 ± 1.897**	0.811 ± 0.065**
(d)	29.632 ± 2.322**	0.863 ± 0.057**	27.533 ± 1.874**	0.813 ± 0.065**
(u)+(d)	29.728 ± 2.349**	0.863 ± 0.056**	27.658 ± 1.905**	0.816 ± 0.065**
(u)+(d)+(t)	29.949 ± 2.391**	0.867 ± 0.057**	27.764 ± 1.906**	0.819 ± 0.064**
(u)+(d)+(t)+(f)	30.007 ± 2.429*	0.869 ± 0.057	27.804 ± 1.925**	0.820 ± 0.065**
(u)+(d)+(t)+(f)+(s)	**30.022 ± 2.428**	**0.869 ± 0.057**	**27.888 ± 1.925**	**0.822 ± 0.064**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.
Abbreviations: PSNR, peak signal-to-noise ratio; SSIM, structural similarity index measure.

by HTC-net improve by 0.024–0.243 dB, indicating that self-supervised pretraining is beneficial for enhancing denoising performance. Table S3 shows the *p*-values for the metrics of networks trained with different methods.

4 | DISCUSSION

In this work, we propose an HTC-net for MR image denoising, which leverages the strengths of Transformer and CNN architectures to enhance the denoising

TABLE 7 Ablation studies on various network configurations on the IXI dataset (mean ± standard deviation): (u) denoising with only the CNN branch, (d) denoising with only the Transformer-CNN branch without Transformer encoder, (t) the Transformer encoder, (f) the RFB, (s) the self-supervised pretraining.

	SNRs = (8, 15)		**SNRs = (5, 8)**	
Method	**PSNR**	**SSIM**	**PSNR**	**SSIM**
(u)	34.709 ± 1.221**	0.967 ± 0.007**	31.216 ± 1.187**	0.930 ± 0.013**
(d)	34.760 ± 1.241**	0.970 ± 0.006**	31.631 ± 1.177**	0.940 ± 0.011**
(u)+(d)	35.175 ± 1.236**	0.971 ± 0.006**	31.767 ± 1.214**	0.941 ± 0.011**
(u)+(d)+(t)	35.232 ± 1.235**	0.972 ± 0.006**	31.980 ± 1.217**	0.945 ± 0.010**
(u)+(d)+(t)+(f)	35.392 ± 1.238**	0.972 ± 0.006**	32.105 ± 1.203**	0.947 ± 0.010**
(u)+(d)+(t)+(f)+(s)	**35.507 ± 1.243**	**0.973 ± 0.005**	**32.275 ± 1.222**	**0.948 ± 0.010**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.
Abbreviations: PSNR, peak signal-to-noise ratio; SNRs, signal-to-noise ratios; SSIM, structural similarity index measure.

TABLE 8 Ablation studies on pretraining strategy on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset.

		SNRs = (8, 15)		**SNRs = (5, 8)**	
Dataset	**Method**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
HP ^{129}Xe MRI dataset	w/o Pretraining	30.007 ± 2.429**	0.869 ± 0.057*	27.804 ± 1.925**	0.820 ± 0.065**
	Pretraining (847 samples)	30.022 ± 2.428	0.869 ± 0.057*	27.888 ± 1.925	0.822 ± 0.064
	Pretraining (1600 samples)	**30.031 ± 2.434**	**0.869 ± 0.057**	**27.894 ± 1.929**	**0.822 ± 0.064**
IXI dataset	w/o Pretraining	35.392 ± 1.238**	0.972 ± 0.006**	32.105 ± 1.203**	0.947 ± 0.010**
	Pretraining (4000 samples)	35.507 ± 1.243**	0.973 ± 0.005**	32.275 ± 1.222**	0.948 ± 0.010**
	Pretraining (8000 samples)	**35.537 ± 1.251**	**0.973 ± 0.005**	**32.348 ± 1.242**	**0.949 ± 0.010**

Note: The best results are shown in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.001, respectively.
Abbreviations: PSNR, peak signal-to-noise ratio; SNRs, signal-to-noise ratios; SSIM, structural similarity index measure.

performance. Additionally, we introduce RFB to effectively aggregate hybrid features at different resolutions. In this way, HTC-net could exploit the comprehensive features to predict accurate noise-free MR images. Moreover, a pretraining strategy based on the self-supervised denoising is utilized to reduce the requirements of paired noisy and noise-free MR images while further improving the performance of the HTC-net. We evaluate the proposed method on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset, demonstrating its superiority over other methods.

Our results demonstrate that long-range information contributes to denoising. In our experiments, it can be observed that leveraging the hybrid network of Transformer and CNN yields superior results compared to classical DL denoising methods using CNN as the basic framework. Specifically, the proposed method outperforms DnCNN by 0.408–0.972 dB in PSNR. This improvement can be attributed to the inherent limitation of CNN in capturing long-range information. Incorporating long-range information aids in minimizing the introduction of false features and maintaining the continuity of anatomical structures, thereby leading to more accurate denoising results (see Figures 4 and 5 and Figure S4). For instance, excellent image denoising methods such as NLM and BM3D are based on the principle of global self-similarity, utilizing the long-range information of the image to improve image denoising performance.[5,9] In our experimental comparisons, it can be observed slightly inferior denoising results on the pulmonary HP ^{129}Xe MRI dataset compared to the IXI dataset. Specifically, our method outperforms DnCNN by 0.408–0.496 dB in PSNR on the pulmonary HP ^{129}Xe MRI dataset, while achieving a higher improvement of 0.694–0.927 dB on the IXI dataset. This disparity can be attributed to the superior image quality and higher spatial resolution of the IXI dataset, which allows for better utilization of long-range information and consequently leads to enhanced denoising performance. In addition, we visualize the shallow and deep feature maps of HTC-net with and without Transformer blocks. It can be noted that HTC-net containing the Transformer blocks provide smoother feature maps and less artifact generation compared to the HTC-net without Transformer blocks (Figure S7). Attention score maps further demonstrate the proposed method effectively leverages long-range information (Figure S8).[39]

After capturing long-range information, effectively integrating local features can further enhance performance. The proposed method demonstrates superior denoising performance compared to SwinIR, as evidenced by higher PSNR and SSIM. SwinIR alternates between convolution and Swin Transformer on a single path, which allows the network to capture long-range information.[40] However, the serial combination of convolution and Swin Transformer may limit the effective aggregation of local and global representations.[41] In our method, the CNN encoder and Transformer encoder work in parallel to extract image features and then effectively aggregate them at different resolutions through the RFBs. This strategy leads to enhanced denoising performance, resulting in a PSNR increase of 0.185–0 .417 dB. K-space fidelity is also important to keep the image content consistent after processing, ensuring that no artifacts are introduced. The proposed method keeps image content consistency by incorporating long-range information and a tailored loss function. This implicitly reduces errors in the k-space data, given that k-space and image-domain data are interconnected via the Fourier transform. Furthermore, explicitly leveraging K-space information is a promising research direction.

Our study also indicates that pretraining is an effective way to improve denoising performance. In natural image restoration, Chen et al. proposed a pretrained Transformer and achieved the state-of-the-art performance.[20] However, this method relies on a large dataset (over 1.1 M images), which is difficult to obtain such a large dataset in MRI. In this work, we propose a pretraining strategy based on self-supervised denoising, which could reduce the requirements of paired noisy and noise-free MR images. Although the self-supervised denoising could not directly recover noise-free MR images from noisy ones due to the non-zero mean, it tends to converge towards outputs without random fluctuations.[24] Therefore, employing self-supervised denoising as pretraining is able to equip the network with denoising capability and facilitate the extraction of MR image features, which is beneficial for subsequent supervised training.[42] Furthermore, increasing the number of noisy images during self-supervised pretraining yields a PSNR increase of 0.024–0 .243 dB. This indicates that by utilizing more noisy MR images during pretraining can decrease the requirements of paired noisy and noise-free images during supervised training, which is particularly suitable for MR image denoising (see Table S4 and Figure S6).

In this work, PSNR and SSIM are both global evaluation metrics, while segmentation involves region of interest (ROI) evaluation. Our method exhibits an effective enhancement in PSNR (0.254–0.927 dB), albeit with comparatively modest improvement in SSIM (0.003–0.016). Regarding ROI evaluation, our method's segmentation results closely resemble those of noise-free images, indicating its capability to provide more accurate quantitative parameters in MR images analysis. In addition, our method demonstrates greater improvements compared to other methods when dealing with lower image quality (at low noise levels: an improvement of 0.254–0.694 dB in PSNR; at high noise levels: an improvement of 0.304–0.927 dB in PSNR). This means that our method is more competitive in scenarios with high levels of noise. The proposed method also shows potential advantages in clinical applications. In tests on under-sampled images, the proposed method demonstrates superior generalization performance compared to other methods, providing high-quality denoising results. Additionally, the improvement of segmentation performance helps to enhance the accuracy of MR image quantitative analysis (Tables 2 and 4, Figures 6 and 9). Moreover, the proposed method exhibits the highest VDP correlation ($R^2 = 0.98$) with noise-free images, contributing to improved diagnosis accuracy (Figures S1 and S2).

However, there are still limitations in this work. First, deeper CNN encoder and Transformer encoder, along with more levels of features aggregation between them, have the potential to improve denoising performance. However, this also increases the risk of overfitting, leading to decreased DL model generalization. Therefore, further research are needed to explore the relationship between performance and network structure during practical applications. Second, the comparisons and evaluations in this study are conducted on the pulmonary HP ^{129}Xe MRI dataset and the IXI dataset. In future, we will incorporate a wider variety of data to evaluate the denoising performance of the method.

5 | CONCLUSIONS

In this work, we propose a novel HTC-net and an effective pretraining strategy for MR image denoising. The HTC-net could efficiently capture long-range information and maintain locality by combining the strengths of both Transformer and CNN. The pretraining strategy, based on self-supervised denoising, further improves the denoising performance and reduce the requirements of paired MR images. Experimental results demonstrate that our method preserves more edges and structures, outperforming the state-of-the-art methods by 0.254–0.927 dB in PSNR, which shows superior performance in MR image denoising. The proposed method effectively enhances the quality of MR images, which could help improve the diagnosis accuracy in clinical.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (2022YFC2410000), National Natural Science Foundation of China (82127802, 21921004), Key Research Program of Frontier Sciences (ZDBS-Y-JSC004), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0540000), Hubei Provincial Key Technology Foundation of China (2021ACA013, 2023BAA021), Natural Science Foundation of Hubei Province (2023AFB1061, 2024AFB284), and China Postdoctoral Science Foundation (2022M713218).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are available within the article and its Supporting Information.

REFERENCES

1. Li Z, Xiao S, Wang C, et al. Complementation-reinforced network for integrated reconstruction and segmentation of pulmonary gas MRI with high acceleration. *Med Phys*. 2024;51(1):378-393. doi:10.1002/mp.16591
2. Zhou Q, Li H, Rao Q, et al. Assessment of pulmonary morphometry using hyperpolarized ^{129}Xe diffusion-weighted MRI with variable-sampling-ratio compressed sensing patterns. *Med Phys*. 2023;50(2):867-878. doi:10.1002/mp.16018
3. Macovski A. Noise in MRI. *Magn Reson Med*. 1996;36(3):494-497. doi:10.1002/mrm.1910360327
4. Mohan J, Krishnaveni V, Guo Y. A survey on the magnetic resonance image denoising methods. *Biomed Signal Process Control*. 2014;9:56-69. doi:10.1016/j.bspc.2013.10.007
5. Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego CA, USA. IEEE; 2005:60-65. doi:10.1109/CVPR.2005.38
6. Manjón JV, Carbonell-Caballero J, Lull JJ, García-Martí G, Martí-Bonmatí L, Robles M. MRI denoising using non-local means. *Med Image Anal*. 2008;12(4):514-523. doi:10.1016/j.media.2008.02.004
7. Manjón JV, Coupé P, Buades A, Collins DL, Robles M. New methods for MRI denoising based on sparseness and self-similarity. *Med Image Anal*. 2012;16(1):18-27. doi:10.1016/j.media.2011.04.003
8. Parthiban L, Subramanian R. Medical image denoising using X-lets. In: *Annual IEEE India Conference*, New Delhi, India. IEEE; 2006:1-6. doi:10.1109/INDCON.2006.302763
9. Dabov K, Foi A, Katkovnik V, Egiazarian K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans Image Process*. 2007;16(8):2080-2095. doi:10.1109/tip.2007.901238
10. Maggioni M, Katkovnik V, Egiazarian K, Foi A. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE Trans Image Process*. 2013;22(1):119-133. doi:10.1109/tip.2012.2210725
11. Xu Y, Han K, Zhou Y, Wu J, Xie X, Xiang W. Deep adaptive blending network for 3D magnetic resonance image denoising. *IEEE J Biomed Health Inform*. 2021;25(9):3321-3331. doi:10.1109/jbhi.2021.3087407
12. Augustin AM, Kesavadas C, Sudeep PV. An improved deep persistent memory network for Rician noise reduction in MR images. *Biomed Signal Process Control*. 2022;77:103736. doi:10.1016/j.bspc.2022.103736
13. 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(7):3142-3155. doi:10.1109/tip.2017.2662206
14. Jiang D, Dou W, Vosters L, Xu X, Sun Y, Tan T. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. *Jpn J Radiol*. 2018;36(9):566-574. doi:10.1007/s11604-018-0758-8
15. You X, Cao N, Lu H, Mao M, Wang W. Denoising of MR images with Rician noise using a wider neural network and noise range division. *Magn Reson Imaging*. 2019;64:154-159. doi:10.1016/j.mri.2019.05.042
16. Ran M, Hu J, Chen Y, et al. Denoising of 3D magnetic resonance images using a residual encoder-decoder Wasserstein generative adversarial network. *Med Image Anal*. 2019;55:165-180. doi:10.1016/j.media.2019.05.001
17. Aetesam H, Maji SK. Noise dependent training for deep parallel ensemble denoising in magnetic resonance images. *Biomed Signal Process Control*. 2021;66:102405. doi:10.1016/j.bspc.2020.102405
18. Liang J, Cao J, Sun G, et al. SwinIR: image restoration using swin transformer. In: *18th IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021. doi:10.1109/ICCVW54120.2021.00210
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *31st Annual Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA. ACM; 2017:4-9. doi:10.48550/arXiv.1706.03762
20. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z. Pre-trained image processing transformer. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021:12294-12305. doi:10.1109/CVPR46437.2021.01212
21. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: a general U-shaped transformer for image restoration. In: *Paper presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA. IEEE; 2022:17662-17672. doi:10.1109/CVPR52688.2022.01716
22. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer; 2021:36-46. doi:10.1007/978-3-030-87193-2_4
23. Lehtinen J, Munkberg J, Hasselgren J, et al. Noise2Noise: learning image restoration without clean data. In: *35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden. PMLR; 2018. doi:10.48550/arXiv.1803.04189
24. Batson J, Royer L. Noise2Self: blind denoising by self-supervision. In: *Paper presented at: 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA. PMLR; 2019. doi:10.48550/arXiv.1901.11365
25. Krull A, Buchholz T-O, Jug F, Soc IC. Noise2void-learning denoising from single noisy images. In: *32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA. IEEE; 2019:2124-2132. doi:10.1109/CVPR.2019.00223
26. Quan Y, Chen M, Pang T, Ji H. Self2Self with dropout: learning self-supervised denoising from single image. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE; 2020:1887-1895. doi:10.1109/CVPR42600.2020.00196
27. Ke RH, Schönlieb CB. Unsupervised image restoration using partially linear denoisers. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(9):5796-5812. doi:10.1109/tpami.2021.3070382
28. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. *Magn Reson Med*. 1995;34(6):910-914. doi:10.1002/mrm.1910340618
29. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv preprint arXiv:201011929*. 2020. doi:10.48550/arXiv.2010.11929
30. Koay CG, Basser PJ. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *J Magn Reson*. 2006;179(2):317-322. doi:10.1016/j.jmr.2006.01.016
31. He M, Zha W, Tan F, Rankine L, Fain S, Driehuys B. A comparison of two hyperpolarized ^{129}Xe MRI ventilation quantification pipelines: the effect of signal to noise ratio. *Acad Radiol*. 2019;26(7):949-959. doi:10.1016/j.acra.2018.08.015
32. Santyr G, Kanhere N, Morgado F, Rayment JH, Ratjen F, Couch MJ. Hyperpolarized gas magnetic resonance imaging of pediatric cystic fibrosis lung disease. *Acad Radiol*. 2019;26(3):344-354. doi:10.1016/j.acra.2018.04.024
33. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA. 2019. doi:10.48550/arXiv.1711.05101

34. Qu X, Zhang W, Guo D, Cai C, Cai S, Chen Z. Iterative thresholding compressed sensing MRI based on contourlet transform. *Inverse Probl Sci Eng*. 2010;18(6):737-758. doi:10.1080/17415977.2010.492509
35. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612. doi:10.1109/tip.2003.819861
36. Guo F, Svenningsen S, Kirby M, et al. Thoracic CT-MRI coregistration for regional pulmonary structure-function measurements of obstructive lung disease. *Med Phys*. 2017;44(2):1718-1733. doi:10.1002/mp.12160
37. Vinurajkumar S, Anandhavelu S. An enhanced fuzzy segmentation framework for extracting white matter from T1-weighted MR images. *Biomed Signal Process Control*. 2022;71:103093. doi:10.1016/j.bspc.2021.103093
38. Chen F, Taviani V, Malkiel I, et al. Variable-density single-shot fast spin-echo MRI with deep learning reconstruction by using variational networks. *Radiology*. 2018;289(2):366-373. doi:10.1148/radiol.2018180445
39. Kim K, Wu B, Dai X, et al. Rethinking the self-attention in vision transformers. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021. doi:10.1109/CVPRW53098.2021.00342
40. Liu Z, Lin YT, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *18th IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:9992-10002. doi:10.1109/ICCV48922.2021.00986
41. Wang Q, Li Z, Zhang S, Chi N, Dai Q. A versatile wavelet-enhanced CNN-transformer for improved fluorescence microscopy image restoration. *Neural Netw*. 2024;170:227-241. doi:10.1016/j.neunet.2023.11.039
42. He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA. IEEE; 2022:15979-15988. doi:10.1109/CVPR52688.2022.01553

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shi S, Wang C, Xiao S, et al. Magnetic resonance image denoising for Rician noise using a novel hybrid transformer-CNN network (HTC-net) and self-supervised pretraining. *Med Phys*. 2025;52:1643–1660. https://doi.org/10.1002/mp.17562