PSDN: PHYSICS-GUIDED SELF-SUPERVISED DUAL-STREAM NETWORK FOR ULTRASOUND IMAGE DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound imaging is widely used in clinical diagnosis but inherently suffers from speckle noise, which degrades image quality and obscures vital anatomical details. Existing deep learning–based denoising methods are severely constrained by the lack of clean ground-truth images, while current self-supervised approaches often neglect acoustic physical mechanisms and rely on single-stream architectures, resulting in frequency aliasing and an inherent trade-off between noise suppression and edge preservation. To address these issues, we propose a Physics-Guided Self-Supervised Dual-Stream Network (**PSDN**) that effectively suppresses speckle noise while preserving anatomical structures and fine details. First, we propose a Physics-Guided Adaptive Frequency Decomposition (**PAFD**) module to alleviate the aliasing between tissue details and speckles using physics-inspired image-statistical priors and dynamically decomposing the noisy image into purified frequency subbands. Second, a Structure-Detail Dual-Stream (**S2DS**) module is proposed to balance macro-structure recovery and micro-detail modeling by processing the decomposed subbands through parallel low- and high-frequency branches. Extensive experimental results across multiple datasets demonstrate that PSDN achieves state-of-the-art denoising performance without requiring clean reference images.

1 INTRODUCTION

Owing to its affordability, non-invasive nature, and real-time imaging capability Jiang et al. (2023), ultrasound imaging plays an important role in disease screening, clinical diagnosis Michailovich & Adam (2005), and image-guided interventional procedures. It reconstructs internal anatomical structures by transmitting acoustic waves and receiving the corresponding echoes Szabo (2013). However, the coherent superposition of echoes scattered by microscopic tissues generates speckle noise, which not only reduces tissue contrast Bottenus et al. (2020); Wu et al. (2022) and obscures subtle structures, but also severely compromises the reliability of clinical diagnosis Huang et al. (2022).

Consequently, ultrasound image denoising has been extensively investigated, with existing methods broadly categorized into three main directions. Early studies mainly relied on classical filtering techniques, including local adaptive filtering Zhang et al. (2015), non-local means (NLM) Zhou et al. (2019), and anisotropic diffusion Yu & Acton (2002); Aja-Fernández & Alberola-López (2006). Although effective to some extent, these methods rely heavily on handcrafted priors and parameters, often blurring boundaries and losing anatomical details. To overcome these limitations, fully supervised deep neural networks, such as DnCNN Zhang et al. (2017) and its variants Gu et al. (2019); Guo et al. (2019), were subsequently introduced and substantially improved denoising performance through their powerful nonlinear representation capabilities. However, because ultrasound speckle noise is intrinsically coupled with tissue echo signals during image formation, obtaining completely noise-free reference images remains impractical in clinical settings. To alleviate this limitation, self-supervised denoising methods have attracted increasing attention in recent years, including deep image prior (DIP) Ulyanov et al. (2018), Noise2Noise (N2N) Lehtinen et al. (2018) and its variants Asgariandehkordi et al. (2025); Jung et al. (2024); Cho et al. (2024), blind-spot networks represented

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

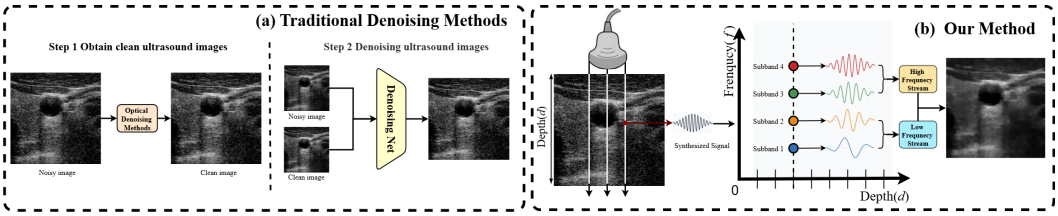

Figure 1: Conceptual comparison between conventional supervised denoising pipelines and the proposed self-supervised PSDN. Conventional supervised approaches generally rely on paired noisy and clean images, whereas PSDN learns to reconstruct a denoised image from a single noisy B-mode ultrasound image through physics-guided frequency decomposition and structure-detail reconstruction.

by Noise2Void (N2V) Krull et al. (2019), and ultrasound-specific self-supervised frameworks Li et al. (2025).

Despite removing the dependence on clean labels, existing self-supervised methods still suffer from two major limitations. First, most approaches model speckle as spatially independent random noise while overlooking its coherent interference mechanism, making it difficult to distinguish structured speckle from genuine tissue information. Second, the strong spatial entanglement between tissue textures and speckle patterns forces conventional single-stream networks to trade off anatomical detail preservation against effective noise suppression. Therefore, an explicit frequency-aware mechanism is needed to separately model tissue structures, fine anatomical details, and noise components across different frequency bands.

To address these limitations, we propose a Physics-Guided Self-Supervised Dual-Stream Network (PSDN). First, we introduce a Physics-Guided Adaptive Frequency Decomposition (PAFD) module, which uses image-derived proxies of acoustic attenuation, local scattering, and tissue structure to spatially modulate a fixed Log-Gabor filter bank. Rather than assuming that each subband exclusively represents either anatomy or noise, we regard the lower-frequency subbands as structure-dominant components and the highest-frequency subband as a speckle-sensitive mixed component that may also contain fine anatomical details. Accordingly, a structure-aware soft weighting strategy is employed to suppress unreliable high-frequency responses while retaining boundary-related information. The resulting subbands are subsequently processed by the Structure-Detail Dual-Stream (S2DS) module, which employs a low-frequency structure branch and a high-frequency detail branch for frequency-specific modeling. Their complementary features are fused to reconstruct the denoised image, achieving a balance between speckle suppression and anatomical detail preservation. In summary, the contributions of this paper are as follows:

- We propose a PAFD module, which leverages acoustic attenuation, local scattering, and tissue-structure priors to generate four adaptively refined frequency subbands with reduced spectral entanglement between anatomical details and speckle noise.
- An S2DS module is developed to suppress speckle noise while preserving anatomical structures by complementarily fusing parallel low-frequency structure and high-frequency detail branches.
- Extensive comparative and ablation experiments on three ultrasound speckle denoising datasets demonstrate the superior performance of PSDN and validate the effectiveness of its key components.

2 RELATED WORK

2.1 TRADITIONAL FILTERING-BASED DENOISING METHODS

Early image denoising approaches predominantly relied on non-learning-based techniques, whose core principles were built upon handcrafted priors and statistical signal modeling. For ultrasound speckle reduction, conventional filtering methods can generally be categorized into three groups: local adaptive filtering Tay et al. (2010); Zhang et al. (2015), non-local means filtering Zhu et al.

(2017); Zhou et al. (2019), and diffusion-based filtering Yu & Acton (2002); Aja-Fernández & Alberola-López (2006); Krissian et al. (2007); Zhou et al. (2014). Although these approaches do not require noise-free ground-truth images, their limited capability for feature representation makes it difficult to effectively disentangle speckle noise from genuine tissue details Lefkimmiatis (2018). This limitation has subsequently motivated the development of deep learning-based ultrasound denoising methods.

2.2 Supervised Learning-Based Denoising Methods

With the rapid development of deep neural networks, the denoising paradigm has shifted toward directly learning an end-to-end mapping from noisy observations to their clean counterparts. Fully supervised models represented by DnCNN Zhang et al. (2017) have substantially outperformed conventional filtering-based methods and have served as the foundation for numerous subsequent architectures Lefkimmiatis (2018); Plötz & Roth (2018); Gu et al. (2019); Guo et al. (2019), continuously advancing image restoration performance. However, these methods rely heavily on paired noisy–clean training samples. In real-world clinical ultrasound acquisition, obtaining standard noise-free reference images is nearly impossible, and the severe scarcity of accurately paired training data substantially limits the practical applicability of supervised ultrasound denoising methods Krull et al. (2019).

2.3 Self-Supervised Learning-Based Denoising Methods

To eliminate the dependence on clean ground-truth images, self-supervised denoising has emerged as an important technical direction for ultrasound image restoration. Existing approaches can be broadly divided into three categories: single-image iterative methods based on deep image priors Ulyanov et al. (2018), dual-image self-supervised methods based on statistically independent observations, such as Noise2Noise Lehtinen et al. (2018) and its variants Asgariandehkordi et al. (2025); Jung et al. (2024); Cho et al. (2024), and single-image methods based on masking or sub-image splitting strategies, including Noise2Void (N2V) Krull et al. (2019), blind-spot network (BSN) architectures Li et al. (2025); Lee et al. (2022); Jang et al. (2023), and Speckle2Self Li et al. (2025), which was specifically developed for ultrasound images. Although existing self-supervised denoising algorithms perform well in general blind denoising scenarios, they suffer obvious drawbacks when applied to clinical ultrasound images. These models fail to incorporate the physical principles of ultrasound imaging, and their unified feature processing frameworks cannot effectively decouple intertwined tissue textures and structured speckle noise, resulting in an inevitable trade-off between speckle suppression and edge preservation, as well as difficulties in separating anatomical information from speckle noise.

3 METHODS

3.1 OVERVIEW

We propose a Physics-Guided Self-Supervised Dual-Stream Network (PSDN) for despeckling real ultrasound images without paired speckle-free reference images. Because ultrasound speckle is coupled with acoustic propagation, tissue scattering, and anatomical structures, PSDN does not treat speckle as independent additive noise. Instead, it uses physics-inspired statistical priors extracted from the input B-mode image to adaptively model different frequency components.

As shown in Fig. 2, PSDN consists of a Physics-Guided Adaptive Frequency Decomposition (PAFD) module and a Structure–Detail Dual-Stream (S2DS) network. PAFD extracts a relative attenuation prior, a local scattering statistics prior, and a tissue structure response prior. Guided jointly by these priors, it decomposes the input image into four semantic subbands ordered from low to high frequency. S2DS then uses a low-frequency structure branch to restore tissue contours and regional intensity distributions, and a high-frequency detail branch to reconstruct tissue boundaries and fine textures. Features from the two branches are fused to produce the despeckled image.

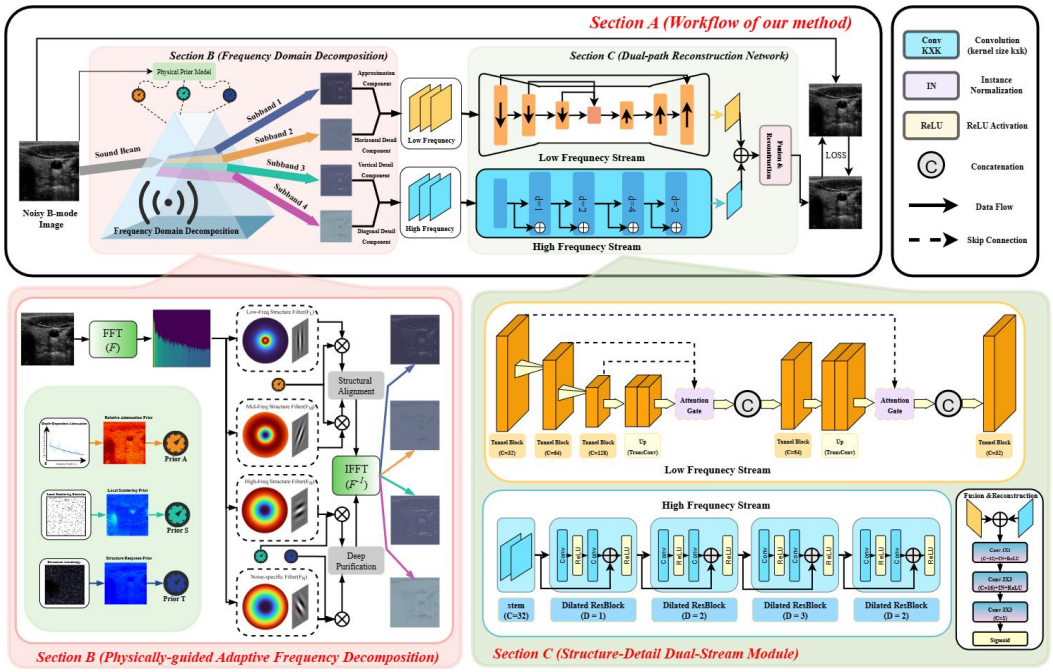

Figure 2: Detailed architecture of the proposed PSDN. Section A presents the complete self-supervised denoising workflow. Section B illustrates the Physics-Guided Adaptive Frequency Decomposition (PAFD) module, and Section C presents the Structure-Detail Dual-Stream (S2DS) reconstruction network. The features produced by the two reconstruction branches are fused to generate the final denoised image.

3.2 Physics-Guided Adaptive Frequency Decomposition

PAFD uses attenuation, scattering, and tissue-structure information to guide the frequency decomposition of ultrasound images. Given an input B-mode image I, we extract three physics-inspired image-statistical priors, denoted by A, S, and T. Taking the attenuation prior as an example, the depth-dependent high-frequency response is fitted by

$$\log \bar{E}_h(z) = \log E_0 - \alpha z, \qquad A(x,z) = \mathrm{Clip}\left[\mathcal{N}\left(1 - \frac{E_h(x,z)}{E_0 \exp(-\alpha z) + \epsilon}\right), 0, 1\right], \tag{1}$$

where $E_h(x,z)$ is the local high-frequency response, $\bar{E}_h(z)$ is its lateral average at depth z, and α is estimated through least-squares fitting. The scattering prior S is estimated from local variance, extrema density, and high-frequency energy ratio, whereas the tissue-structure prior T integrates gradient, Laplacian, and structure-tensor responses. These quantities are image-derived statistical proxies rather than quantitative estimates of physical acoustic parameters. Their complete definitions and implementation procedures are provided in the Supplementary Material.

PAFD decomposes I using four Log-Gabor filters. The radial response of the k-th filter is defined as

$$G_k(r) = \exp\left[-\frac{\log^2(r/f_k)}{2\log^2(\sigma_k)}\right], \tag{2}$$

where f_k and σ_k control the center frequency and bandwidth, respectively. The resulting subband is spatially modulated by the three priors:

$$B_k^{(0)} = \mathrm{Clip}\left[(1 - q_k A)(1 - q_k S)(0.5 + 0.5T), 0, 1\right] \odot \mathcal{F}^{-1}\left[\mathcal{F}(I) \odot G_k\right], \tag{3}$$

where $q_k = (k-1)/(K-1)$ controls the prior strength across frequency bands.

Previous multiscale studies on ultrasound despeckling have shown that the low-frequency approximation band retains most of the global image content, whereas directional high-frequency subbands

preserve horizontal, vertical, and diagonal edge information (Vimalraj et al., 2016). Meanwhile, speckle fluctuations also produce pronounced responses in multiscale high-frequency representations. Therefore, wavelet- and Gabor-based despeckling methods generally distinguish edge-related responses from speckle-related fluctuations instead of indiscriminately removing all high-frequency coefficients (Yue et al., 2006; Zhang et al., 2014; Khor et al., 2022). Based on these observations, we interpret B_1–B_3 as structure-dominant bands at progressively finer scales, while B_4 is treated as a speckle-sensitive high-frequency mixed band that may contain both irregular speckle fluctuations and genuine anatomical boundaries.

To prevent the direct removal of genuine anatomical details in B_4, we employ a structure-preserving soft gate:

$$E_N = \mathrm{ReLU}(A \odot S - 0.3T), \qquad \widehat{B}_k = B_k^{(0)} \odot \exp(-2.5 q_k E_N). \tag{4}$$

Thus, high-frequency responses are strongly suppressed only in regions with strong attenuation and scattering but weak structural evidence. The four refined subbands $\{\widehat{B}_k\}_{k=1}^4$ are subsequently fed into S2DS for structure-detail reconstruction.

3.3 Structure–Detail Dual-Stream network

Different frequency components have distinct spatial characteristics. Low-frequency subbands mainly contain large-scale tissue contours and regional intensity variations, whereas high-frequency subbands contain fine boundaries, tissue textures, and speckle responses. Processing all frequency components with a single network may lead to oversmoothed structures or residual high-frequency speckle. We therefore design the Structure–Detail Dual-Stream (S2DS) network to model low-frequency structural information and high-frequency mixed information separately.

The low-frequency structure branch concatenates the first two subbands along the channel dimension:

$$X_L = \mathrm{Cat}(\widehat{B}_1, \widehat{B}_2). \tag{5}$$

An encoder–decoder extracts large-scale tissue contours, regional intensity distributions, and spatial context. During decoding, attention-guided skip connections transfer useful shallow structural information while suppressing residual speckle. The resulting structural features are

$$F_L = \mathcal{F}_L(X_L), \tag{6}$$

where $\mathcal{F}_L(\cdot)$ denotes the low-frequency structure branch.

The high-frequency detail branch takes the third and fourth subbands as input:

$$X_H = \mathrm{Cat}(\widehat{B}_3, \widehat{B}_4). \tag{7}$$

This branch maintains the original spatial resolution throughout and applies residual blocks with dilation rates of 1, 2, 3, and 2 in sequence:

$$F_H = \mathcal{R}_2 \circ \mathcal{R}_3 \circ \mathcal{R}_2 \circ \mathcal{R}_1 \left[\mathrm{Stem}(X_H)\right], \tag{8}$$

where $\mathcal{R}_d$ denotes a residual block with dilation rate d. The resulting multiscale receptive fields allow the network to consider both local responses and longer-range structural continuity when distinguishing isolated speckle fluctuations from continuous tissue boundaries. Notably, $\widehat{B}_4$ is treated as a speckle-sensitive high-frequency mixed band rather than a pure noise band. It remains an input to the detail branch so that genuine edges and fine textures within it can contribute to reconstruction.

Finally, features from the two branches are added element-wise and passed to a fusion and reconstruction module:

$$\widehat{I} = \mathrm{Sigmoid}\left[\mathcal{C}_r(F_L + F_H)\right], \tag{9}$$

where $\mathcal{C}_r(\cdot)$ comprises 1×1 and 3×3 convolutions for channel interaction and spatial reconstruction, respectively. By combining complementary low-frequency structural and high-frequency detail features, S2DS suppresses speckle while preserving tissue contours, boundaries, and fine-grained textures.

Table 1: Quantitative comparisons between our proposed algorithm and seven mainstream state-of-the-art methods (1 traditional filtering method, 1 supervised denoising method, 5 self-supervised denoising methods) are conducted on three datasets.

Method	Carotid				Clinical Kidney				Calf Muscle			
	ENL↑	CNR↑	Hom↑	PS↑	ENL↑	CNR↑	Hom↑	PS↑	ENL↑	CNR↑	Hom↑	PS↑
Traditional Filter	9.42	1.03	0.198	2.91	8.85	0.98	0.203	2.83	8.26	0.91	0.190	2.77
N2V	8.76	1.21	0.215	3.14	8.32	1.16	0.221	3.09	8.51	1.09	0.208	3.03
N2S	9.31	1.24	0.227	3.21	8.91	1.19	0.232	3.15	9.05	1.12	0.219	3.10
Neighbor2Neighbor	9.88	1.27	0.213	3.26	9.27	1.22	0.218	3.21	9.46	1.16	0.211	3.16
ZS-N2N	10.12	1.29	0.214	3.31	9.58	1.24	0.224	3.27	9.71	1.18	0.216	3.22
Speckle2Self	15.74	1.61	**0.725**	4.02	14.62	1.54	**0.738**	3.94	14.93	1.47	**0.701**	3.88
PSDN (Ours)	**17.36**	**1.78**	0.578	**4.28**	**16.28**	**1.71**	0.571	**4.21**	**16.67**	**1.64**	0.546	**4.17**

3.4 Self-Supervised Loss Function

Because real ultrasound images generally lack precisely aligned speckle-free references, we construct a self-supervised pseudo-target from the frequency subbands generated by PAFD. The highest-frequency subband, $\widehat{B}_4$, contains both speckle and genuine anatomical details. We therefore retain it with a soft weight determined by structural reliability:

$$R_4 = T \odot \exp(-\eta E_N), \tag{10}$$

$$Y_p = \mathrm{Clip}\left(\widehat{B}_1 + \widehat{B}_2 + \widehat{B}_3 + R_4 \odot \widehat{B}_4,\, 0,\, 1\right), \tag{11}$$

where E_N is the local noise response estimated by PAFD and $\eta = 2.5$. High-frequency responses in $\widehat{B}_4$ are retained where there is strong evidence of tissue structure, while responses without structural support are downweighted. Thus, the highest-frequency subband is not treated as pure noise.

The overall loss is defined as

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{stv}}\mathcal{L}_{\mathrm{stv}} + \lambda_{\mathrm{ene}}\mathcal{L}_{\mathrm{ene}}, \tag{12}$$

where the reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}} = \left\| \widehat{I} - \mathrm{SG}(Y_p) \right\|_1. \tag{13}$$

Here, $\mathrm{SG}(\cdot)$ denotes the stop-gradient operation. $\mathcal{L}_{\mathrm{rec}}$ provides the primary self-supervised training signal; $\mathcal{L}_{\mathrm{stv}}$ uses the scattering prior to suppress local fluctuations in speckle-active regions; and $\mathcal{L}_{\mathrm{ene}}$ constrains the overall response scale before and after frequency decomposition. We set $\lambda_{\mathrm{rec}} = 1$, $\lambda_{\mathrm{stv}} = 0.3$, and $\lambda_{\mathrm{ene}} = 0.1$. Complete definitions of the loss terms and parameter analyses are provided in the Supplementary Material.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets

We utilize three ultrasound datasets for experiments, the carotid ultrasound dataset Li et al. (2025), which consists of real clinical B-mode carotid images with a resolution of 512×512 without synthetic samples; the clinical kidney ultrasound dataset Song et al. (2022), collected from two hospitals and containing 391 transverse and longitudinal B-mode images of healthy kidneys; the calf muscle ultrasound dataset Michard et al. (2021), integrated from two public gastrocnemius datasets with a total of 813 raw images, all resized to a uniform resolution of 512×512 in our experiments.

Evaluation Metrics. Because strictly paired speckle-free reference images are unavailable for the three real-world ultrasound datasets, we do not report full-reference metrics such as PSNR, SSIM, or LPIPS. Instead, we evaluate despeckling performance using the equivalent number of looks (ENL), contrast-to-noise ratio (CNR), homogeneity (Hom), and blinded physician score (PS). ENL measures speckle suppression within homogeneous regions, while CNR quantifies the contrast between

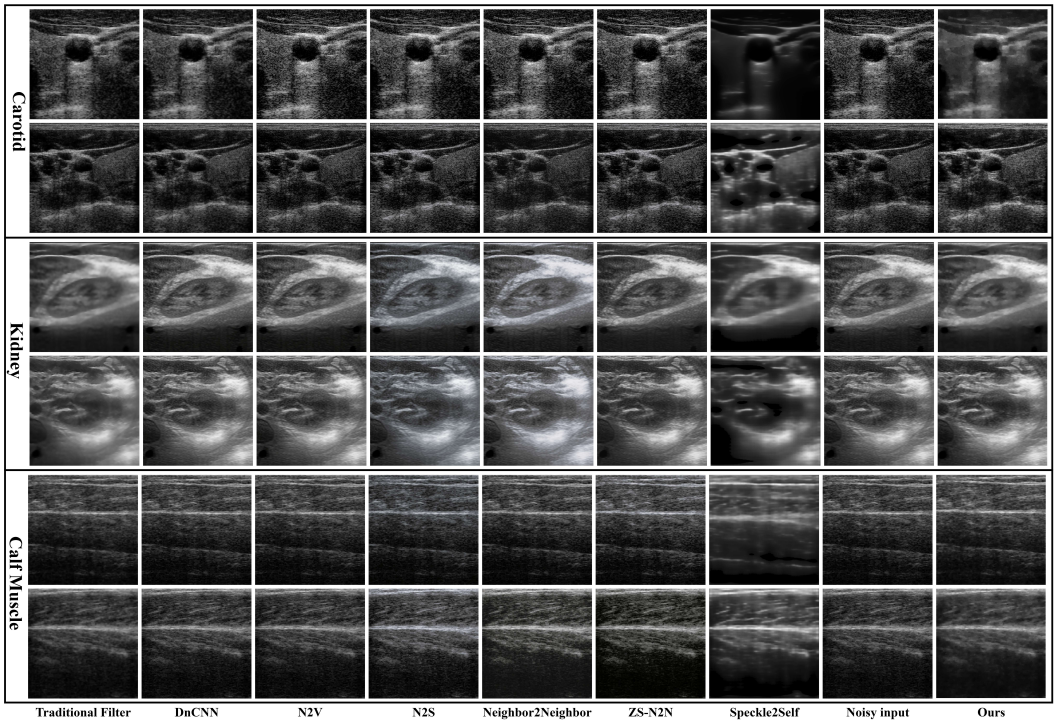

Figure 3: Qualitative comparison results on three datasets.

selected tissue and background regions. Hom characterizes regional intensity uniformity; however, because excessive smoothing may also increase Hom, this metric is interpreted jointly with CNR, qualitative structural preservation, and physician assessment rather than as an independent indicator of overall image quality. For PS, three ultrasound physicians independently evaluated the anonymized and randomly ordered results using a five-point scale, considering speckle suppression, boundary preservation, texture naturalness, and clinical acceptability. Higher values indicate better performance for ENL, CNR, and PS, whereas Hom is used specifically to characterize regional uniformity. The metric definitions, region-of-interest selection protocol, and physician scoring criteria are provided in the Supplementary Material.

Implementation Details The entire training pipeline is built on the PyTorch framework with a batch size of 4. Data augmentation strategies including horizontal flip, vertical flip, as well as shift-scale-rotation transformation are adopted, with augmentation probabilities set to 0.5, 0.5 and 0.4 respectively. The Adam optimizer is utilized for training over a total of 6000 iterations. The initial learning rate is set to 0.001 and gradually decayed to 0.0001.

4.2 COMPARISON WITH STATE-OF-THE-ART

Comparison Methods We compare PSDN with seven representative methods, including a traditional diffusion-based filter Zhou et al. (2014), and five self-supervised methods: N2V Krull et al. (2019), N2S Batson & Royer (2019), Neighbor2Neighbor Huang et al. (2021), ZS-N2N Mansour & Heckel (2023), and Speckle2Self Li et al. (2025). All methods are evaluated on the same test images using identical preprocessing and the same pre-specified regions of interest. Official implementations and recommended settings are used whenever available. Since strictly paired speckle-free reference images are unavailable, no full-reference metrics, including PSNR, SSIM, or LPIPS, are reported. Instead, ENL, CNR, Hom, and blinded physician scores are used to evaluate speckle suppression, tissue contrast, regional uniformity, and clinical acceptability, respectively.

Quantitative Comparison. Table 1 presents the quantitative comparison results on the carotid, clinical kidney, and calf muscle datasets. PSDN achieves the highest ENL, CNR, and physician score (PS) on all three datasets, with ENL/CNR/Hom/PS values of 17.36/1.78/0.578/4.28,

16.28/1.71/0.571/4.21, and 16.67/1.64/0.546/4.17, respectively. Compared with Speckle2Self, which shows the strongest overall performance among the competing methods, PSDN improves ENL by 1.62, 1.66, and 1.74, CNR by 0.17, 0.17, and 0.17, and PS by 0.26, 0.27, and 0.29 on the three datasets, respectively. Speckle2Self obtains higher Hom values of 0.725, 0.738, and 0.701, compared with 0.578, 0.571, and 0.546 for PSDN. This result indicates that Speckle2Self produces more homogeneous local regions, which may be associated with stronger smoothing of speckle-like fluctuations. However, Hom primarily measures regional intensity uniformity and does not directly quantify the preservation of anatomical boundaries or diagnostically relevant fine textures. Therefore, a higher Hom value alone should not be interpreted as superior clinical image quality. The higher CNR and blinded physician scores achieved by PSDN suggest that it provides a more favorable balance among speckle suppression, tissue-contrast enhancement, anatomical-detail preservation, and clinical acceptability.

Qualitative Comparison. Fig. 3 presents the visual comparison between the proposed method and seven SOTA denoising algorithms across three ultrasound datasets. The results demonstrate that our method produces sharp boundaries and natural tissue textures while effectively suppressing speckle artifacts. In contrast, most competing methods suffer from over-smoothing, edge blurring, or residual speckle. In particular, Speckle2Self strongly suppresses speckle fluctuations but produces an overly smooth appearance, removing part of the fine anatomical textures and subtle boundaries. Such aggressive despeckling is not necessarily preferable for clinical ultrasound because speckle-related echo patterns may contain diagnostically relevant tissue information. For the carotid artery dataset, the structural prior embedded in the PAFD module accurately locates vascular boundaries, while the dual-stream architecture prevents edge over-smoothing, yielding clear vascular contours and uniform tissue grayscale. For the kidney dataset, the priors of ultrasound attenuation and scattering adapt to inter-tissue echo differences, and frequency-split processing balances contour integrity and internal detail retention, achieving distinct renal cortical-medullary layering and natural texture transitions. For the calf muscle dataset, spectral-domain frequency decomposition separates muscle fiber signals from speckle noise, which, paired with the high-frequency branch for precise texture recognition, generates muscle fiber textures with minimal residual noise and high coherence.

4.3 Ablation Study

Ablation Study of Each Component Table 2 evaluates the contributions of PAFD and S2DS on the carotid dataset. The baseline obtains an ENL of 6.42, a CNR of 1.03, and a Hom of 0.184. Introducing PAFD alone improves these values to 13.85, 1.48, and 0.421, respectively, demonstrating its effectiveness in physics-guided frequency decomposition and speckle suppression. Using S2DS alone increases ENL, CNR, and Hom to 10.27, 1.36, and 0.472, confirming its contribution to structural and detailed feature reconstruction. Combining both modules achieves the best performance, with an ENL of 17.36, a CNR of 1.78, and a Hom of 0.578. Compared with the baseline, the complete model improves the three metrics by 10.94, 0.75, and 0.394, respectively. These results demonstrate that PAFD and S2DS provide complementary benefits in speckle suppression, tissue-contrast enhancement, and anatomical-structure preservation.

Table 2: Ablation study of components on the Carotid Ultrasound Dataset.

PAFD	S2DS	ENL↑	CNR↑	Hom↑
×	×	6.42	1.03	0.184
✓	×	13.85	1.48	0.421
×	✓	10.27	1.36	0.472
✓	✓	**17.36**	**1.78**	**0.578**

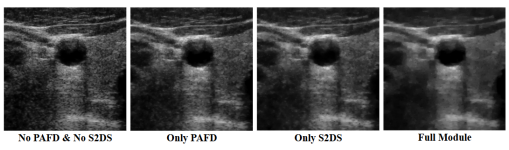

Figure 4: Qualitative comparison of the component-ablation variants on the carotid ultrasound dataset.

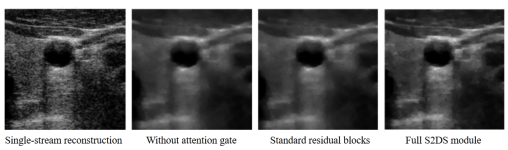

Figure 5: Qualitative comparison of different S2DS variants on the carotid ultrasound dataset.

Ablation Study of S2DS. Table 3 compares different variants of the S2DS module. The single-stream reconstruction model obtains the lowest performance, indicating that uniformly processing frequency components limits structure–detail modeling. Removing the attention gate decreases ENL, CNR, and Hom to 15.62, 1.64, and 0.519, respectively, because residual speckle responses may be transferred through the skip connections. Replacing the dilated residual blocks with standard residual blocks also reduces the three metrics to 16.31, 1.70, and 0.546, demonstrating the importance of multi-scale contextual modeling. The complete S2DS module achieves the best ENL, CNR, and Hom of 17.36, 1.78, and 0.578, respectively, verifying that the dual-stream architecture, attention gates, and dilated residual blocks jointly improve speckle suppression and anatomical-detail preservation.

Table 3: Ablation of different S2DS variants on the carotid ultrasound dataset.

S2DS Variants	ENL↑	CNR↑	Hom↑
Single-stream reconstruction	14.08	1.52	0.458
Without attention gate	15.62	1.64	0.519
Standard residual blocks	16.31	1.70	0.546
Full S2DS module	**17.36**	**1.78**	**0.578**

Table 4: Sensitivity to the number of frequency bands on the carotid ultrasound dataset.

Number of Bands K	ENL↑	CNR↑	Hom↑
3	15.62	1.62	0.523
4 (Ours)	**17.36**	**1.78**	**0.578**
5	16.89	1.73	0.561
6	16.44	1.69	0.548

Sensitivity Analysis of Frequency Decomposition. We investigate the robustness of PAFD with respect to the number of frequency bands and the Log-Gabor filter parameters. As shown in Table 4, using three bands produces relatively low performance because the coarse spectral partition cannot sufficiently distinguish global structures, fine anatomical details, and speckle-sensitive responses. The four-band configuration achieves the best ENL, CNR, and Hom of 17.36, 1.78, and 0.578, respectively. Increasing the number to five or six does not provide further improvements, which may be attributed to increased spectral overlap and redundancy between adjacent subbands. Therefore, four frequency bands are adopted in PSDN.

Table 5: Sensitivity analysis of the Log-Gabor parameters on the carotid ultrasound dataset.

Metric	Center Frequency Scale					Bandwidth Scale				
	0.90×	0.95×	1.00×	1.05×	1.10×	0.80×	0.90×	1.00×	1.10×	1.20×
ENL↑	16.71	17.08	**17.36**	17.14	16.83	16.61	17.02	**17.36**	17.11	16.74
CNR↑	1.71	1.75	**1.78**	1.75	1.72	1.69	1.74	**1.78**	1.75	1.71
Hom↑	0.552	0.568	**0.578**	0.570	0.557	0.545	0.563	**0.578**	0.567	0.551

We further perturb the center frequencies and bandwidths of the Log-Gabor filters around their default values. As reported in Table 5, moderate changes of $\pm 5\%$ result in only small performance variations, whereas larger deviations cause gradual performance degradation. The default configuration achieves the best overall performance, but the relatively stable results within the tested range indicate that PAFD does not depend on a narrowly selected Log-Gabor configuration.

5 Conclusion

In this paper, we propose PSDN to suppress speckle noise while preserving tissue boundaries and fine-grained anatomical structures. Ultrasound images inevitably suffer from strong signal-dependent speckle, which blurs tiny tissue features and interferes with clinical diagnosis. First, a PAFD module is introduced to integrate physics-inspired image-statistical priors for the adaptive decomposition and refinement of image frequency components, thereby alleviating spectral entanglement between anatomical details and speckle noise. This frequency separation strategy avoids the common defect of simultaneous loss of texture details in traditional filtering algorithms. Second, we propose an S2DS module to model low- and high-frequency information separately and perform complementary feature fusion for image reconstruction. Low-frequency components retain global tissue contours, while high-frequency components capture subtle anatomical textures for accurate feature complementation.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing and improving the clarity and readability of the manuscript. No generative AI tools were used to generate experimental results, data, or scientific claims. All AI-assisted text was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work

ETHICS STATEMENT

This study uses anonymized ultrasound images for research purposes. No personally identifiable information is included. The proposed method is not intended to replace professional clinical judgment.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training objectives, experimental settings, and evaluation protocols in the main paper and supplementary material. Additional implementation details and experimental results are provided to facilitate reproducibility. Code and relevant resources will be released upon acceptance, subject to applicable data and licensing restrictions.

REFERENCES

Santiago Aja-Fernández and Carlos Alberola-López. On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering. *IEEE Transactions on Image Processing*, 15(9):2694–2701, 2006.

Hojat Asgariandehkordi, Mostafa Sharifzadeh, and Hassan Rivaz. Lightweight physics-informed zero-shot ultrasound plane wave denoising. *arXiv preprint arXiv:2506.21499*, 2025.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pp. 524–533. PMLR, 2019.

Nick Bottenus, Melissa LeFevre, Jayne Cleve, Anna Lisa Crowley, and Gregg Trahey. Resolution and speckle reduction in cardiac imaging. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 68(4):1131–1143, 2020.

Hyunwoo Cho, Seongjun Park, Jinbum Kang, and Yangmo Yoo. Deep coherence learning: An unsupervised deep beamformer for high quality single plane wave imaging in medical ultrasound. *Ultrasonics*, 143:107408, 2024.

Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2511–2520, 2019.

Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1712–1722, 2019.

Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14781–14790, 2021.

Zixun Huang, Rui Zhao, Frank HF Leung, Sunetra Banerjee, Timothy Tin-Yan Lee, De Yang, Daniel PK Lun, Kin-Man Lam, Yong-Ping Zheng, and Sai Ho Ling. Joint spine segmentation and noise removal from ultrasound volume projection images with selective feature sharing. *IEEE transactions on medical imaging*, 41(7):1610–1624, 2022.

Yeong Il Jang, Keuntek Lee, Gu Yong Park, Seyun Kim, and Nam Ik Cho. Self-supervised image denoising with downsampled invariance loss and conditional blind-spot network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12196–12205, 2023.

Zhongliang Jiang, Septimiu E Salcudean, and Nassir Navab. Robotic ultrasound imaging: State-of-the-art and future perspectives. *Medical image analysis*, 89:102878, 2023.

Dongkyu Jung, Myeongkyun Kang, Sang Hyun Park, Nizar Guezzi, and Jaesok Yu. Unsupervised speckle noise reduction technique for clinical ultrasound imaging. *Ultrasonography*, 43(5):327–344, 2024.

Hee Guan Khor, Guochen Ning, Xinran Zhang, and Hongen Liao. Ultrasound speckle reduction using wavelet-based generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3080–3091, 2022. doi: 10.1109/JBHI.2022.3144628.

Karl Krissian, Carl-Fredrik Westin, Ron Kikinis, and Kirby G Vosburgh. Oriented speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 16(5):1412–1424, 2007.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17725–17734, 2022.

Stamatios Lefkimmiatis. Universal denoising networks: a novel cnn architecture for image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3204–3213, 2018.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

Xuesong Li, Nassir Navab, and Zhongliang Jiang. Speckle2self: Self-supervised ultrasound speckle reduction without clean data. *Medical image analysis*, pp. 103755, 2025.

Youssef Mansour and Reinhard Heckel. Zero-shot noise2noise: Efficient image denoising without any data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14018–14027, 2023.

Oleg V Michailovich and Dan Adam. A novel approach to the 2-d blind deconvolution problem in medical ultrasound. *IEEE transactions on medical imaging*, 24(1):86–104, 2005.

Hugo Michard, Bertrand Luvison, Quoc-Cuong Pham, Antonio J Morales-Artacho, and Gaël Guilhem. Aw-net: Automatic muscle structure analysis on b-mode ultrasound images for injury prevention. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–9, 2021.

Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural information processing systems*, 31, 2018.

Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022.

Thomas L Szabo. *Diagnostic ultrasound imaging: inside out*. Academic press, 2013.

Peter C Tay, Christopher D Garson, Scott T Acton, and John A Hossack. Ultrasound despeckling for contrast enhancement. *IEEE Transactions on Image Processing*, 19(7):1847–1860, 2010.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

C Vimalraj, S Esakkirajan, Thangaraj Veerakumar, and P Sreevidya. Direction sensitive wavelet packet for despeckling of ultrasound images. *IET Computer Vision*, 10(7):746–757, 2016.

Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022.

Yongjian Yu and Scott T Acton. Speckle reducing anisotropic diffusion. *IEEE Transactions on image processing*, 11(11):1260–1270, 2002.

Yong Yue, Mihai M. Croitoru, Akhil Bidani, Joseph B. Zwischenberger, and John W. Clark. Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images. *IEEE Transactions on Medical Imaging*, 25(3):297–311, 2006. doi: 10.1109/TMI.2005.862737.

Ju Zhang, Guangkuo Lin, Lili Wu, Chen Wang, and Yun Cheng. Wavelet and fast bilateral filter based de-speckling method for medical ultrasound images. *Biomedical Signal Processing and Control*, 18:1–10, 2015.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Qi Zhang, Hong Han, Chunhong Ji, Jinhua Yu, Yuanyuan Wang, and Wenping Wang. Gabor-based anisotropic diffusion for speckle noise reduction in medical ultrasonography. *Journal of the Optical Society of America A*, 31(6):1273–1283, 2014.

Yingyue Zhou, Hongbin Zang, Su Xu, Hongsen He, Jinzheng Lu, and Hongdao Fang. An iterative speckle filtering algorithm for ultrasound images based on bayesian nonlocal means filter model. *Biomedical signal processing and control*, 48:104–117, 2019.

Zhenyu Zhou, Zhichang Guo, Gang Dong, Jiebao Sun, Dazhi Zhang, and Boying Wu. A doubly degenerate diffusion model based on the gray level indicator for multiplicative noise removal. *IEEE Transactions on Image Processing*, 24(1):249–260, 2014.

Lei Zhu, Chi-Wing Fu, Michael S Brown, and Pheng-Ann Heng. A non-local low-rank framework for ultrasound speckle reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5650–5658, 2017.

A Complete Definitions of Physics-Inspired Priors

The attenuation, scattering, and tissue-structure priors are calculated directly from the input B-mode ultrasound image without requiring radio-frequency data, acquisition parameters, or manual annotations. These priors should be interpreted as physics-inspired image-statistical proxies rather than quantitative estimates of acoustic tissue properties.

For an arbitrary response map X, the min–max normalization is defined as

$$\mathcal{N}(X) = \frac{X - \min(X)}{\max(X) - \min(X) + \varepsilon}, \tag{14}$$

where $\varepsilon = 10^{-8}$ prevents numerical instability.

A.1 Attenuation Prior

Given an input image I, we first calculate its local high-frequency response using Sobel operators:

$$E_h(x, z) = \sqrt{(\nabla_x I(x, z))^2 + (\nabla_z I(x, z))^2 + \varepsilon}, \tag{15}$$

where x and z denote the lateral position and imaging depth, respectively. The laterally averaged response at depth z is

$$\overline{E}_h(z) = \frac{1}{W} \sum_{x=1}^{W} E_h(x, z). \tag{16}$$

An exponential depth-decay model is fitted by least squares in the logarithmic domain:

$$\log\left(\overline{E}_h(z)+\varepsilon\right)=\log E_0-\alpha z, \tag{17}$$

where E_0 and α are the fitted initial response and decay coefficient. The depth-dependent expected response is

$$E_{\exp}(z)=E_0\exp(-\alpha z). \tag{18}$$

The attenuation prior is then defined as

$$A(x,z)=\mathrm{Clip}\left[\mathcal{N}\left(1-\frac{E_h(x,z)}{E_{\exp}(z)+\varepsilon}\right),0,1\right]. \tag{19}$$

A large value of A indicates that the observed high-frequency response is weaker than its depth-dependent expectation.

A.2 Scattering Prior

The scattering prior combines local variance, local-maximum density, and the local high-frequency spectral ratio. For a 7×7 neighborhood Ω_p centered at pixel p, the local variance is

$$V(p)=\frac{1}{|\Omega_p|}\sum_{q\in\Omega_p}I^2(q)-\left(\frac{1}{|\Omega_p|}\sum_{q\in\Omega_p}I(q)\right)^2. \tag{20}$$

Let $M(p)$ indicate whether $I(p)$ is a local maximum within a 3×3 neighborhood. The local-maximum density is calculated as

$$D(p)=\frac{1}{|\Omega_p|}\sum_{q\in\Omega_p}M(q). \tag{21}$$

For each local image patch I_{Ω_p}, the high-frequency spectral ratio is defined as

$$R_{\mathrm{HF}}(p)=\frac{\sum_{(u,v)\in\Omega_{\mathrm{HF}}}\left|\mathcal{F}(I_{\Omega_p})(u,v)\right|}{\sum_{u,v}\left|\mathcal{F}(I_{\Omega_p})(u,v)\right|+\varepsilon}, \tag{22}$$

where Ω_{HF} denotes the predefined high-frequency region of the local spectrum.

The three normalized responses are combined as

$$S_0=0.4\mathcal{N}(V)+0.3\mathcal{N}(D)+0.3\mathcal{N}(R_{\mathrm{HF}}), \tag{23}$$

$$S=\mathrm{Sigmoid}\left[5(S_0-0.5)\right]. \tag{24}$$

A high value of S indicates strong local intensity fluctuations, dense extrema, and a relatively large high-frequency response.

A.3 Tissue-Structure Prior

The tissue-structure prior combines gradient magnitude, Laplacian response, and structure-tensor anisotropy:

$$G=\sqrt{(\nabla_x I)^2+(\nabla_z I)^2+\varepsilon}, \tag{25}$$

$$L=\left|\nabla^2 I\right|. \tag{26}$$

The local structure tensor is calculated as

$$\mathbf{J}=\begin{bmatrix}K_\sigma*(\nabla_x I)^2 & K_\sigma*(\nabla_x I\nabla_z I)\\ K_\sigma*(\nabla_x I\nabla_z I) & K_\sigma*(\nabla_z I)^2\end{bmatrix}, \tag{27}$$

where K_σ denotes a 5×5 Gaussian kernel with $\sigma=1$, and $*$ denotes convolution.

Let λ_1 and λ_2 be the eigenvalues of $\mathbf{J}$, where $\lambda_1 \geq \lambda_2$. The anisotropy response is defined as

$$C = \frac{\lambda_1}{\lambda_2 + \varepsilon}. \tag{28}$$

The final tissue-structure prior is

$$T_0 = 0.5\mathcal{N}(G) + 0.3\mathcal{N}(L) + 0.2\mathcal{N}(C), \tag{29}$$

$$T = \mathrm{Sigmoid}\left[5(T_0 - 0.5)\right]. \tag{30}$$

A large value of T indicates a strong and directionally consistent response that is more likely to correspond to a genuine anatomical boundary.

B Implementation Details of PAFD

B.1 Log-Gabor Filter Bank

PAFD employs four radial Log-Gabor filters with normalized center frequencies

$$(f_1, f_2, f_3, f_4) = (0.05, 0.15, 0.30, 0.45) \tag{31}$$

and bandwidth parameters

$$(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.08, 0.12, 0.15, 0.10). \tag{32}$$

For the normalized radial frequency

$$r(u, v) = \sqrt{f_u^2 + f_v^2}, \tag{33}$$

the k-th Log-Gabor filter is defined as

$$G_k(r) = \exp\left[-\frac{\log^2(r/f_k)}{2\log^2(\sigma_k)}\right]. \tag{34}$$

At the frequency origin, $r(0, 0)$ is replaced with ε, and the DC response is explicitly set to

$$G_k(0, 0) = 0. \tag{35}$$

The Log-Gabor filters remain fixed during training and do not contain learnable parameters.

B.2 Physics-Guided Spatial Modulation

The frequency-dependent prior strength is defined as

$$q_k = \frac{k-1}{K-1}, \qquad K = 4. \tag{36}$$

The spatial modulation map for the k-th subband is

$$W_k = \mathrm{Clip}\left[(1 - q_k A)(1 - q_k S)(0.5 + 0.5T), 0, 1\right]. \tag{37}$$

The initial subband is obtained by

$$B_k^{(0)} = W_k \odot \mathcal{F}^{-1}\left[\mathcal{F}(I) \odot G_k\right], \tag{38}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and inverse Fourier transform, respectively.

We further define a local noise-evidence map:

$$E_N = \mathrm{ReLU}(A \odot S - 0.3T). \tag{39}$$

The structure-preserving soft-gated subbands are calculated as

$$\widehat{B}_k = B_k^{(0)} \odot \exp(-2.5 q_k E_N). \tag{40}$$

This gate suppresses high-frequency responses more strongly in regions with simultaneous attenuation and scattering evidence but weak structural evidence. When T is large, the suppression is reduced to preserve genuine boundaries and fine tissue textures.

B.3 Interpretation of the Four Subbands

The four subbands represent dominant image characteristics at different frequency scales rather than strictly disjoint semantic categories. Specifically, $\widehat{B}_1$ mainly contains global intensity distributions and macroscopic tissue structures; $\widehat{B}_2$ represents regional tissue transitions; $\widehat{B}_3$ emphasizes fine-scale anatomical structures; and $\widehat{B}_4$ contains a mixture of subtle boundaries, tissue textures, and speckle-sensitive fluctuations. This interpretation is consistent with previous ultrasound despeckling studies showing that directional high-frequency subbands contain anatomically meaningful edge responses in addition to speckle-related fluctuations (**?**Yue et al., 2006; **?**; Khor et al., 2022). Therefore, the semantic descriptions of the four subbands indicate their dominant characteristics rather than mutually exclusive signal categories. In particular, $\hat{B}_4$ is not discarded as a pure noise component; instead, its spatial responses are selectively retained according to the tissue-structure evidence encoded by T.

Therefore, $\widehat{B}_4$ is treated as a speckle-sensitive high-frequency mixed band rather than a pure noise component.

B.4 Soft Retention of High-Frequency Structures

To avoid discarding genuine anatomical information in $\widehat{B}_4$, we define a spatially varying structural-reliability weight:

$$R_4 = T \odot \exp(-2.5E_N). \tag{41}$$

The pseudo-clean target is constructed as

$$Y_p = \mathrm{Clip}\left(\widehat{B}_1 + \widehat{B}_2 + \widehat{B}_3 + R_4 \odot \widehat{B}_4, 0, 1\right). \tag{42}$$

Consequently, high-frequency responses supported by strong tissue-structure evidence are retained, whereas spatially irregular responses without structural support are downweighted.

C Architectural Details of S2DS

The Structure–Detail Dual-Stream network contains a low-frequency structure branch and a high-frequency detail branch. The base number of feature channels is set to 32.

C.1 Low-Frequency Structure Branch

The input to the structure branch is

$$X_L = \mathrm{Cat}(\widehat{B}_1, \widehat{B}_2). \tag{43}$$

The branch adopts an encoder–decoder architecture with two stride-2 convolutional downsampling stages and two transposed-convolution upsampling stages. It models macroscopic tissue contours, regional intensity distributions, and long-range spatial context.

An attention gate is applied to the first skip connection:

$$Q = \phi_x(F_{\mathrm{enc}}) + \phi_g(F_{\mathrm{dec}}), \tag{44}$$

$$M_{\mathrm{att}} = \mathrm{Sigmoid}\left[\psi\left(\mathrm{ReLU}(Q)\right)\right], \tag{45}$$

$$F_{\mathrm{skip}} = M_{\mathrm{att}} \odot F_{\mathrm{enc}}, \tag{46}$$

where ϕ_x, ϕ_g, and ψ are learnable 1×1 convolutions. The attention gate selectively transfers useful shallow structural information while reducing residual speckle responses in the skip connection.

C.2 High-Frequency Detail Branch

The input to the detail branch is

$$X_H = \mathrm{Cat}(\widehat{B}_3, \widehat{B}_4). \tag{47}$$

Table 6: Architectural configuration of the S2DS network.

Module	Input Ch.	Output Ch.	Kernel	Stride/Dilation	Resolution
Structure stem	2	32	3×3	1	$H \times W$
Structure encoder 1	32	C_1	3×3	2	$\frac{H}{2} \times \frac{W}{2}$
Structure encoder 2	C_1	C_2	3×3	2	$\frac{H}{4} \times \frac{W}{4}$
Structure decoder 1	C_2	C_1	4×4	2	$\frac{H}{2} \times \frac{W}{2}$
Structure decoder 2	C_1	32	4×4	2	$H \times W$
Detail stem	2	32	3×3	1	$H \times W$
Residual block 1	32	32	3×3	$d = 1$	$H \times W$
Residual block 2	32	32	3×3	$d = 2$	$H \times W$
Residual block 3	32	32	3×3	$d = 3$	$H \times W$
Residual block 4	32	32	3×3	$d = 2$	$H \times W$
Fusion convolution	32	32	1×1	1	$H \times W$
Output convolution	32	1	3×3	1	$H \times W$

The detail branch maintains the original spatial resolution throughout the reconstruction process. It contains four dilated residual blocks with dilation rates

$$(1, 2, 3, 2). \tag{48}$$

For an input feature map F, a residual block with dilation rate d is expressed as

$$R_d(F) = F + \mathcal{C}_{d,2}\left[\delta\left(\mathcal{C}_{d,1}(F)\right)\right], \tag{49}$$

where $\mathcal{C}_{d,1}$ and $\mathcal{C}_{d,2}$ denote 3×3 dilated convolutions, and δ denotes the nonlinear activation function.

The detail feature is obtained by

$$F_H = R_2 \circ R_3 \circ R_2 \circ R_1\left[\mathrm{Stem}(X_H)\right]. \tag{50}$$

C.3 Dual-Stream Fusion

Let F_L and F_H denote the outputs of the structure and detail branches, respectively. They are fused by element-wise addition:

$$F_{\mathrm{fuse}} = F_L + F_H. \tag{51}$$

The denoised image is reconstructed using a 1×1 convolution for channel interaction, followed by a 3×3 convolution for spatial reconstruction:

$$\widehat{I} = \mathrm{Sigmoid}\left[\mathcal{C}_{3\times3}\left(\mathcal{C}_{1\times1}(F_{\mathrm{fuse}})\right)\right]. \tag{52}$$

D Complete Definitions of Training Objectives

D.1 Pseudo-Target Reconstruction Loss

The pseudo-target reconstruction loss is defined as

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\widehat{I}_{i,j} - \mathrm{SG}(Y_{p,i,j})\right|, \tag{53}$$

where $\mathrm{SG}(\cdot)$ denotes the stop-gradient operation. Therefore, the pseudo-target construction does not receive gradients from the reconstruction loss.

D.2 Scattering-Weighted Total-Variation Loss

To suppress spatially irregular fluctuations in speckle-active regions, the total-variation loss is weighted using the scattering prior:

$$\mathcal{L}_{\mathrm{stv}} = \frac{1}{HW} \sum_{i,j} (1 + \beta S_{i,j}) \left(\left| \nabla_x \widehat{I}_{i,j} \right| + \left| \nabla_y \widehat{I}_{i,j} \right| \right), \tag{54}$$

where β controls the contribution of the scattering prior. In our implementation, $\beta = 1$.

D.3 Relative Subband-Response Regularization

The relative subband-response regularization is defined as

$$\mathcal{L}_{\mathrm{ene}} = \frac{\left| \|I\|_2^2 - \sum_{k=1}^{4} \|\widehat{B}_k\|_2^2 \right|}{\|I\|_2^2 + \varepsilon}. \tag{55}$$

This term limits excessive changes in the overall response magnitude during frequency decomposition. Since the Log-Gabor filters have overlapping spectral supports, it is used as a response-scale regularizer rather than a strict spectral-energy conservation constraint.

D.4 Overall Objective

The complete self-supervised training objective is

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{rec}} \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{stv}} \mathcal{L}_{\mathrm{stv}} + \lambda_{\mathrm{ene}} \mathcal{L}_{\mathrm{ene}}, \tag{56}$$

where

$$\lambda_{\mathrm{rec}} = 1, \qquad \lambda_{\mathrm{stv}} = 0.3, \qquad \lambda_{\mathrm{ene}} = 0.1. \tag{57}$$

All three loss terms are calculated from the noisy input image and its derived subbands. No paired speckle-free reference image is used during training.

E Dataset Specifications

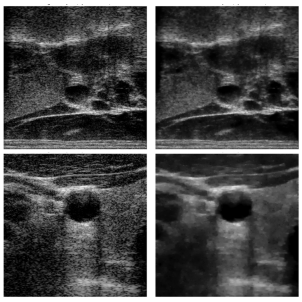

Figure 6: Representative carotid inputs and PSDN outputs.

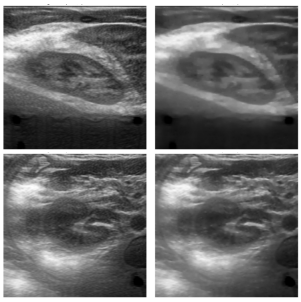

Figure 7: Representative kidney inputs and PSDN outputs.

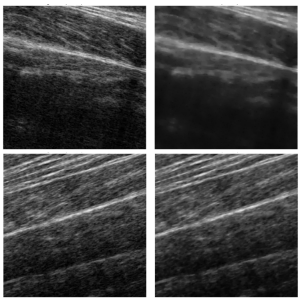

Figure 8: Representative calf-muscle inputs and PSDN outputs.

We evaluate PSDN on three real-world B-mode ultrasound datasets covering different anatomical structures. All images are resized to 512×512 and normalized to $[0, 1]$. PSDN is trained without paired speckle-free targets, and all methods are evaluated using the same test images, preprocessing, and pre-specified regions of interest. Because strictly paired clean references are unavailable, PSNR, SSIM, and LPIPS are not reported.

Carotid Ultrasound Dataset (Li et al., 2025). The carotid ultrasound dataset contains real two-dimensional B-mode carotid ultrasound images acquired using a portable ultrasound device. The

images include representative vascular-wall and plaque structures and are used to evaluate whether the proposed method can suppress perivascular speckle while preserving vessel boundaries and fine anatomical structures. Representative noisy inputs and corresponding denoising results are shown in Figure 6.

Clinical Kidney Ultrasound Dataset (Song et al., 2022). The clinical kidney ultrasound dataset was collected from two independent medical institutions and contains 391 transverse and longitudinal B-mode ultrasound images of healthy kidneys without cysts or tumors. This dataset covers renal cortical and medullary structures and is used to evaluate despeckling performance on parenchymal organs. Representative noisy inputs and corresponding denoising results are shown in Figure 7.

Calf Muscle Ultrasound Dataset (Michard et al., 2021). The calf muscle ultrasound dataset integrates two public gastrocnemius ultrasound datasets, denoted as A2 and B2, and contains a total of 813 original images. All images are resized to 512×512 during preprocessing. The fine-grained muscle-fiber patterns in this dataset are used to evaluate the ability of PSDN to suppress speckle while preserving subtle anatomical textures. Representative noisy inputs and corresponding denoising results are shown in Figure 8.

F Evaluation Protocol

The three datasets do not provide strictly paired speckle-free references; therefore, PSNR, SSIM, and LPIPS are not reported. All methods are evaluated on the same images and pre-specified regions of interest (ROIs) using equivalent number of looks (ENL), contrast-to-noise ratio (CNR), homogeneity (Hom), and blinded physician score (PS). ENL measures fluctuation suppression, CNR measures tissue-background contrast, and Hom characterizes local uniformity. Because smoothing may also increase ENL or Hom, all metrics are interpreted jointly with structural appearance and physician assessment.

F.1 Equivalent Number of Looks

For a homogeneous ROI Ω_h, ENL is calculated as

$$\mathrm{ENL}(\Omega_h) = \frac{\mu_h^2}{\sigma_h^2 + \epsilon}, \tag{58}$$

where μ_h and σ_h^2 are the mean and variance within Ω_h, respectively, and $\epsilon = 10^{-8}$. Higher ENL indicates weaker local fluctuations. Values are averaged across the selected ROIs and then across test images.

F.2 Contrast-to-Noise Ratio

For a target-tissue ROI Ω_t and its corresponding background ROI Ω_b, CNR is defined as

$$\mathrm{CNR}(\Omega_t, \Omega_b) = \frac{|\mu_t - \mu_b|}{\sqrt{\sigma_t^2 + \sigma_b^2 + \epsilon}}, \tag{59}$$

where (μ_t, σ_t^2) and (μ_b, σ_b^2) are the mean and variance of the target and background ROIs. Higher CNR indicates stronger contrast relative to local fluctuations. Multiple ROI pairs are averaged within each image and then across the test set.

F.3 Homogeneity

Homogeneity is calculated from the normalized gray-level co-occurrence matrix (GLCM). Let $P_d^\theta(i, j)$ denote the normalized probability of observing gray levels i and j at spatial distance d and direction θ. Hom is defined as

$$\mathrm{Hom}_d^\theta = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \frac{P_d^\theta(i,j)}{1+(i-j)^2}, \tag{60}$$

where L is the number of quantized gray levels. Values are averaged over the evaluated directions, homogeneous ROIs, and test images. Higher Hom indicates stronger local uniformity, but may also reflect over-smoothing and is not interpreted independently as superior clinical quality.

F.4 ROI Selection Protocol

ROIs were selected on the original noisy images before any method outputs were inspected, and their coordinates were transferred unchanged to all results. ENL and Hom use fixed-size square ROIs from homogeneous tissue without strong boundaries, shadows, artifacts, lesions, or image borders. CNR uses equal-sized target-tissue and adjacent-background ROIs. No method-specific ROI adjustment was permitted.

F.5 Blinded Physician Evaluation

Three ultrasound physicians independently scored anonymized and randomly ordered outputs without method information. The five-point scale jointly considers speckle suppression, boundary preservation, texture naturalness, and clinical acceptability: 1 denotes severe artifacts or distortion, 3 acceptable quality with moderate residual speckle or smoothing, and 5 excellent suppression with sharp boundaries and natural texture; scores 2 and 4 represent intermediate quality. For method m,

$$\mathrm{PS}_m = \frac{1}{N_{\mathrm{img}} N_{\mathrm{reader}}} \sum_{n=1}^{N_{\mathrm{img}}} \sum_{r=1}^{N_{\mathrm{reader}}} s_{m,n,r}, \tag{61}$$

where $s_{m,n,r}$ is the score assigned by reader r to image n, N_{img} is the number of test images, and $N_{\mathrm{reader}} = 3$. ENL, CNR, Hom, and PS are interpreted jointly to assess the balance between speckle suppression, contrast, anatomical-detail preservation, and clinical acceptability.