Frequency-Aware Post-Training Quantization for Medical Image Denoising

Uni Ko[1], Yumi Lee[1], Juneyoung Park[2], and Jang-Hwan Choi[1]✉

[1] Ewha Womans University, Seoul, Korea
{uniko,eumi,choij}@ewha.ac.kr
[2] OptAI Inc., Korea
jyoung.park@opt-ai.kr

Abstract. Deep learning denoisers achieve remarkable restoration quality, yet their high computational cost limits deployment in resource-constrained settings. While post-training quantization (PTQ) enables efficient deployment, conventional PTQ methods degrade fidelity under ultra-low precision, causing frequency-selective degradation in reconstructed outputs. Observing that quantization errors concentrate in high-frequency subbands, we propose a dual-stage PTQ framework based on stationary wavelet transform (SWT) that explicitly penalizes subband-wise quantization distortions. Extensive experiments demonstrate consistent improvements over existing methods across diverse architectures. Our method effectively balances fidelity with substantial reductions in model size, supporting deployment-oriented medical image denoising. The code is available at https://github.com/unikohoho/Quantization.

Keywords: Medical Image Denoising · Post-Training Quantization · Model Compression.

1 Introduction

Low-dose imaging protocols, including CT and X-ray, are clinically essential for reducing radiation exposure [1]. However, the resulting noise and artifacts can degrade image quality and obscure fine anatomical details [2, 8]. Recent deep learning-based denoising models—ranging from convolutional neural networks (CNNs) [3] to transformer and hybrid architectures [4–7]—achieve high restoration quality but incur substantial computational and memory costs, limiting deployment in resource-constrained medical imaging environments.

Model quantization reduces model size and improves inference efficiency by mapping floating-point weights and activations to low-bit representations. Post-training quantization (PTQ) is a practical solution for deployment, as it does not require access to full training data and avoids costly end-to-end quantization-aware training [13, 18]. However, existing PTQ methods, largely developed for natural-image tasks such as super resolution [14–17], suffer noticeable fidelity degradation when directly applied to medical image denoising, especially under

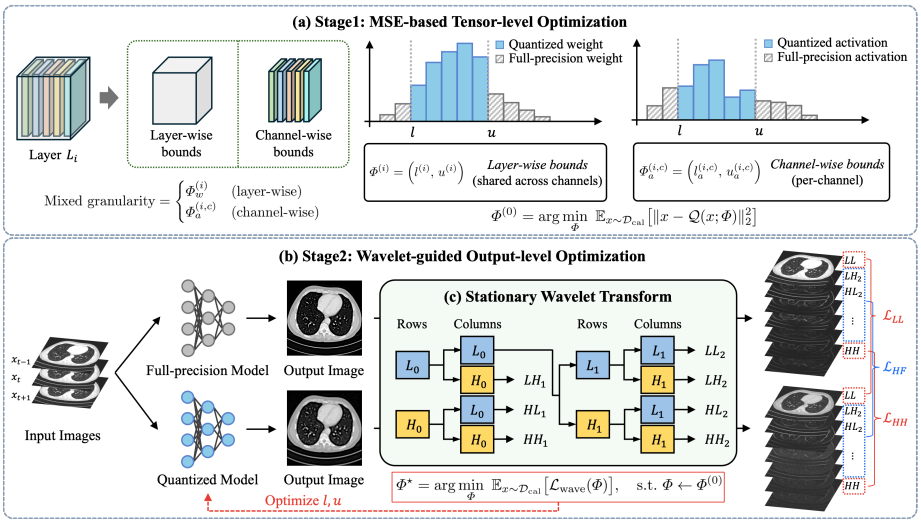

Fig. 1. Overview of the proposed Frequency-Aware PTQ Framework. (a) Tensor-level calibration with mixed-granularity initialization; (b) Wavelet-guided output-level optimization; and (c) SWT decomposition framework.

ultra-low precision ($\leq$4-bit). Recent studies in medical imaging further emphasize the growing need for efficient and deployable models across diverse network architectures [19, 20]. In this context, robust ultra-low-bit PTQ provides a practical route toward reducing the computational and memory burden of medical imaging models without costly retraining.

Directly adopting recent PTQ methods for natural-image restoration [15–17] is insufficient for medical denoising for two main reasons. **First**, recent PTQ methods primarily minimize spatial-domain discrepancies between the quantized and full-precision models. However, quantization distortion is disproportionately concentrated in high-frequency subbands, especially under ultra-low precision. **Second**, existing methods often design quantizers based on specific statistical priors. Such assumptions do not generalize across architectures, as activation distributions vary significantly between backbones.

To address this limitation, we propose the first frequency-aware PTQ framework for medical image denoising (Fig. 1). Leveraging the stationary wavelet transform (SWT), we introduce a dual-stage optimization strategy that directly suppresses subband-level distortions. Our main contributions are:

1. **SWT Analysis of PTQ Degradation:** We show that quantization in medical denoising disproportionately degrades high-frequency subbands, revealing a structural limitation of conventional spatial-domain calibration.
2. **Architecture-Agnostic Frequency-Aware PTQ Framework:** We propose a novel PTQ strategy that directly suppresses frequency-selective dis-

tortion and generalizes across diverse denoising backbones, including CNN, ViT, and hybrid architectures.
3. **High Fidelity with Practical Compression:** Extensive experiments demonstrate that our method consistently outperforms existing PTQ baselines, achieving up to **5.54 dB** PSNR improvement while delivering up to **7.40**× model compression, validated under ONNX Runtime setting.

2 Method

Uniform affine quantization. We adopt standard uniform affine quantization for both weights and activations, as widely used in PTQ for deployment settings [21]. Given tensor $\mathbf{X}$ and bit-width b, clipping bounds (l, u) determine scale $s = \frac{u-l}{2^b-1}$ and zero-point $z = \left\lfloor -\frac{l}{s} \right\rceil$, where $\lfloor \cdot \rceil$ denotes rounding-to-nearest. The quantize–dequantize simulator is formulated as:

$$\mathbf{Q} = \mathrm{clip}\left(\left\lfloor \frac{\mathbf{X}}{s} \right\rceil + z,\ 0,\ 2^b - 1 \right), \qquad \mathbf{X}_Q = s\,(\mathbf{Q} - z), \tag{1}$$

Stationary Wavelet Transform To localize frequency-selective distortion while preserving spatial alignment, we adopt the stationary wavelet transform (SWT). Unlike the discrete wavelet transform (DWT), SWT removes downsampling and upsamples the low-pass (h) and high-pass (g) filters at each decomposition level j via zero-insertion (Eq. (2)). This preserves the original spatial resolution and provides shift-invariant subband representations [22, 23].

$$h_{j+1}[k] = h_j[k] \uparrow 2, \qquad g_{j+1}[k] = g_j[k] \uparrow 2 \tag{2}$$

2.1 Revisiting the Spatial Bottleneck in Conventional PTQ

The Low-Frequency Bias of Spatial Objectives Recent PTQ strategies commonly rely on spatial-domain losses to align quantized output $\hat{y}$ with full-precision output y. The spatial objective is formulated as:

$$\mathcal{L}_{\text{spatial}} = \|\hat{y} - y\|_p^p, \quad p \in \{1, 2\}. \tag{3}$$

When $p = 2$, Parseval's theorem yields

$$\|\hat{y} - y\|_2^2 = \|\mathcal{F}(\hat{y} - y)\|_2^2 = \sum_{\omega} |E(\omega)|^2, \tag{4}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform and $E(\omega)$ is the spectral error at frequency ω, indicating that spatial L2 loss minimizes total spectral error without differentiating frequency bands [24]. Since medical images exhibit power-law spectral decay ($P(f) \propto 1/f^\beta$) [25, 26] with dominant low-frequency energy, errors in low-frequency bands tend to contribute strongly to the overall objective. As a result, optimization driven by spatial L2 loss implicitly prioritizes low-frequency consistency, while high-frequency distortions tend to be under-penalized.

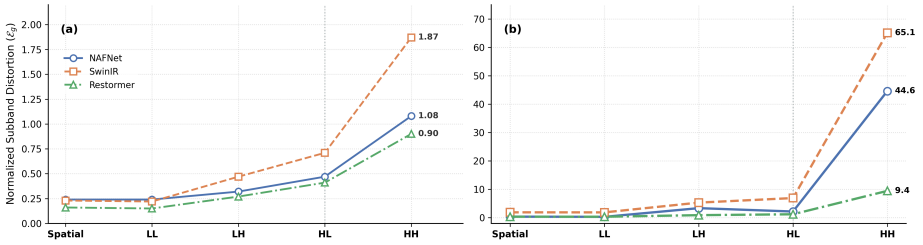

Fig. 2. Subband-wise distortion under 2-bit MinMax PTQ. Results are shown for three denoising architectures (NAFNet, SwinIR, Restormer) on (a) CT and (b) fluoroscopy datasets. For multi-level subbands, we report the level-wise average.

Stationary Wavelet Transform Analysis Using 2-level Haar SWT [27], a reconstruction $y \in \mathbb{R}^{H\times W}$ is decomposed into subbands $\{LL_J, LH_j, HL_j, HH_j\}_{j=1}^{J}$, where $S_g(\cdot)$ denotes extraction of subband g. We apply normalization $\mathcal{N}(\cdot)$ to standardize each subband to zero-mean and unit-variance using statistics estimated from full-precision outputs. Let y^{fp} and y^{q} denote full-precision and quantized reconstructions, respectively. We quantify per-subband distortion $\mathcal{E}_g$ as the mean absolute difference between normalized coefficients:

$$\mathcal{E}_g = \mathbb{E}\left[\left|\mathcal{N}(S_g(y^q)) - \mathcal{N}(S_g(y^{fp}))\right|\right], \tag{5}$$

Across architectures and datasets, the diagonal HH band exhibits the largest distortion (Fig. 2), indicating frequency-selective degradation under ultra-low precision. This observation motivates our frequency-aware refinement strategy.

2.2 Frequency-Aware PTQ Framework

Mixed-granularity initialization. To ensure robustness under ultra-low precision across diverse backbones, we adopt layer-wise quantization for weights and selective channel-wise quantization for activations. Activation statistics vary substantially across network architectures; selective channel-wise activation quantization accommodates inter-channel heterogeneity while avoiding uniform channel-wise treatment for all activation layers. All bounds are initialized from calibration samples and refined through the proposed PTQ framework. Throughout, backbone weights remain frozen and only clipping bounds Φ are optimized.

Tensor-Level Coarse Calibration Given the affine quantizer in Eq. (1), we first optimize clipping bounds by minimizing tensor reconstruction error for each quantized weight/activation tensor. Let x denote a floating-point tensor and $\hat{x} = \mathcal{Q}(x;\Phi)$ its quantized counterpart. We optimize bounds by:

$$\Phi^{(0)} = \arg\min_{\Phi}\ \mathbb{E}_{x\sim\mathcal{D}_{\text{cal}}}\left[\|x - \hat{x}\|_2^2\right], \tag{6}$$

Gradients are propagated through rounding and clipping using a straight-through estimator (STE) [28]. This provides a stable initialization of the clipping bounds

by aligning quantized weights and activations with their full-precision counterparts before frequency-aware output refinement.

Output-Level Wavelet-Guided Fine Optimization Starting from $\Phi^{(0)}$, we refine clipping bounds by explicitly penalizing frequency-selective distortion at the output level. Motivated by the HH-dominant distortion pattern (Sec. 2.1), we group subbands into low-frequency (LL), axis-aligned high-frequency (LH, HL), and diagonal high-frequency (HH) sets. Let $\mathcal{G}_{LL}$, $\mathcal{G}_{HF}$, and $\mathcal{G}_{HH}$ denote the corresponding subband sets, defining the wavelet-guided objective as:

$$\mathcal{L}_{\text{wave}}(\Phi) = \sum_{g \in \mathcal{G}_{LL}} \lambda_{LL}\, \mathcal{E}_g + \sum_{g \in \mathcal{G}_{HF}} \lambda_{HF}\, \mathcal{E}_g + \sum_{g \in \mathcal{G}_{HH}} \lambda_{HH}\, \mathcal{E}_g, \tag{7}$$

We refine the clipping bounds by:

$$\Phi^\star = \arg\min_{\Phi} \mathbb{E}_{x \sim \mathcal{D}_{\text{cal}}} \big[\mathcal{L}_{\text{wave}}(\Phi)\big], \quad \text{initialized at } \Phi^{(0)}. \tag{8}$$

Band-specific weights $\lambda_{LL}, \lambda_{HF}, \lambda_{HH}$ are fixed empirically, with robustness examined in Sec. 3.3. This refinement directly suppresses subband-level distortion.

3 Experiments and Results

3.1 Experimental Setup

Experiments were conducted on the Mayo Clinic Low Dose CT Grand Challenge dataset [9], which provides paired full-dose and quarter-dose CT data, and a fluoroscopic X-ray sequence dataset acquired using a dynamic phantom [10]. For full-precision training, we used 3,994 Mayo frames and 2,335 fluoroscopy frames. Quantitative evaluation was performed on 1,946 Mayo and 1,003 fluoroscopy test frames. Full-precision training and PTQ calibration were performed on NVIDIA A100 GPUs using PyTorch 2.8.0. For deployment evaluation, quantized models were exported to ONNX and executed with ONNX Runtime 1.24.2 using the CUDAExecutionProvider on an NVIDIA RTX GPU with CUDA 12.8.

Backbone Architectures. To validate architecture-agnostic applicability, we evaluate three representative denoising backbones: NAFNet [11], SwinIR [12], and Restormer [5]. These models cover CNN-based, ViT-based, and hybrid convolution-attention architectures.

Implementation Details. To incorporate temporal context, each model receives a 3-frame triplet $\{t-1, t, t+1\}$ stacked along the channel dimension. The calibration set follows the same configuration, where up to 10 triplets per patient are randomly sampled from the training set ($|\mathcal{D}_{\text{cal}}|$ =70 and 120 triplets for CT and fluoroscopy, respectively).

Table 1. Quantitative comparison across datasets. Both weights and activations are quantized to the same bit-width. PSNR/SSIM are reported for 2-, 4-, and 8-bit PTQ across three denoising backbones and two datasets.

Dataset		Mayo CT						Fluoroscopy					
Method	Bit	Restormer		NAFNet		SwinIR		Restormer		NAFNet		SwinIR	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Low-dose	–	33.0887	0.8066	33.0887	0.8066	33.0887	0.8066	37.4166	0.9473	37.4166	0.9473	37.4166	0.9473
FP	32	38.8496	0.9265	38.7638	0.9259	38.7120	0.9253	45.8604	0.9820	43.7798	0.9819	45.5081	0.9811
MinMax	8	36.6356	0.8967	36.5744	0.9010	35.4438	0.8734	44.7881	0.9771	**45.8129**	0.9790	44.2544	0.9730
Percentile	8	38.8039	0.9260	38.7338	0.9254	38.6496	0.9243	45.2726	0.9791	43.8083	0.9809	44.3472	0.9734
MSE	8	36.6357	0.8918	36.5634	0.8932	35.4462	0.8737	44.6665	0.9786	45.7959	0.9810	44.3352	0.9729
PTQ4SR	8	38.8002	0.9255	38.7362	0.9255	38.6443	0.9244	45.7089	0.9808	44.2977	0.9812	44.6859	0.9754
BRECQ	8	37.4005	0.9010	37.4005	0.9010	36.0347	0.8775	44.1665	0.9768	44.4106	0.9787	42.7686	0.9666
2DQuant	8	**38.8201**	0.9258	38.7268	0.9255	38.6421	0.9243	45.7664	0.9811	44.6593	0.9813	44.7392	0.9760
Ours	8	38.8159	**0.9261**	**38.7391**	**0.9256**	**38.6773**	**0.9247**	**45.8269**	**0.9816**	45.2457	**0.9815**	**45.4372**	**0.9807**
MinMax	4	29.6403	0.6005	31.7304	0.7474	32.8654	0.7803	33.6113	0.7687	39.3984	0.9207	29.4347	0.5686
Percentile	4	33.7367	0.8072	37.1671	0.9020	33.2541	0.7783	34.8786	0.8159	41.6331	0.9585	27.1252	0.4306
MSE	4	31.7706	0.7255	33.9812	0.8452	32.6107	0.7596	35.0175	0.9310	38.8354	0.9502	31.5069	0.6723
PTQ4SR	4	34.7275	0.8339	37.4744	0.9082	34.2090	0.8165	41.8363	0.9558	42.8780	0.9706	33.5424	0.7433
BRECQ	4	28.0560	0.5031	31.3385	0.7291	31.3443	0.7226	33.1133	0.7224	35.3940	0.8831	30.4170	0.5658
2DQuant	4	38.0323	0.9167	37.8690	0.9149	37.3021	0.9046	42.8047	0.9705	43.8308	0.9755	39.1118	0.9268
Ours	4	**38.4772**	**0.9229**	**38.1805**	**0.9194**	**38.2002**	**0.9176**	**45.3979**	**0.9797**	**44.4273**	**0.9800**	**44.6516**	**0.9759**
MinMax	2	22.0130	0.7299	17.0378	0.6570	17.9051	0.5472	30.0645	0.9184	27.8654	0.5678	14.8173	0.2871
Percentile	2	16.7759	0.1289	33.1357	0.8080	19.6473	0.1911	13.7468	0.4244	28.6350	0.4726	17.4700	0.1855
MSE	2	29.0069	0.6279	27.7923	0.6370	27.2760	0.7063	23.0319	0.7840	33.5567	0.8671	32.9559	0.7073
PTQ4SR	2	33.0840	0.8063	33.1438	0.8084	28.1592	0.6907	35.0098	0.8730	38.7631	0.9394	37.7231	0.9473
BRECQ	2	28.8306	0.7120	25.4159	0.6193	18.8435	0.4207	27.1026	0.3769	27.1026	0.3769	20.9078	0.2514
2DQuant	2	35.5613	0.8748	35.3894	0.8817	35.5097	0.8876	39.0887	0.9473	40.1942	0.9504	37.2473	0.9473
Ours	2	**37.3524**	**0.9109**	**36.0774**	**0.8906**	**36.4013**	**0.8949**	**42.3603**	**0.9691**	**42.5026**	**0.9730**	**41.8806**	**0.9585**

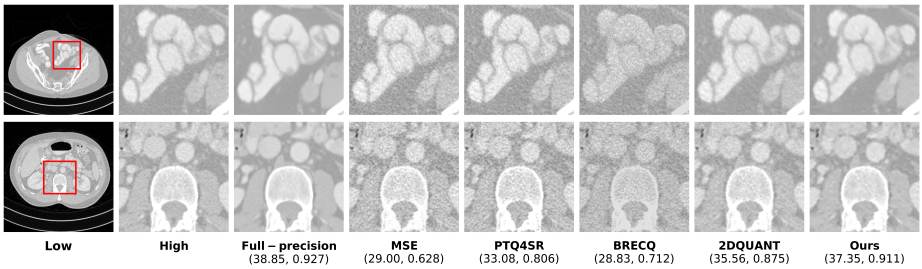

Fig. 3. Qualitative comparison on Mayo CT (2-bit, Restormer backbone). Visual results and corresponding PSNR and SSIM values for 2-bit PTQ.

We evaluate PTQ at $b \in \{2, 4, 8\}$ bits for both weights and activations. Clipping bounds (Φ) are initialized using MinMax. Stage 1 optimizes Φ for 100 iterations using Adam (lr=10^{-3}), and Stage 2 further optimizes for 4,000 iterations using Adam (lr=10^{-4}), both with cosine annealing decay. We use 2-level Haar SWT with subband weights $\lambda_{LL} = 0.2$, $\lambda_{HF} = 0.3$, and $\lambda_{HH} = 0.5$.

3.2 Results

Comparison with PTQ Baselines. We compare against standard clipping methods (MinMax, Percentile, and MSE-based clipping) and strong PTQ baselines, including BRECQ [18], PTQ4SR [15], and 2DQuant [17]. For fair comparison, all baseline methods were implemented following their official implementations and recommended configurations.

Table 2. ONNX-exported model-size reduction. Comparison of model size and compression ratio for the quantized models.

Model		NAFNet		SwinIR		Restormer	
Method	Bit	Size(MB) ↓	Comp. ↑	Size(MB) ↓	Comp. ↑	Size(MB) ↓	Comp. ↑
FP32 (Baseline)	32	65.53	1.00×	2.90	1.00×	101.18	1.00×
Ours	8	16.96	3.86×	0.83	3.49×	26.55	3.81×
Ours	4	8.86	7.40×	0.48	6.04×	14.10	7.17×

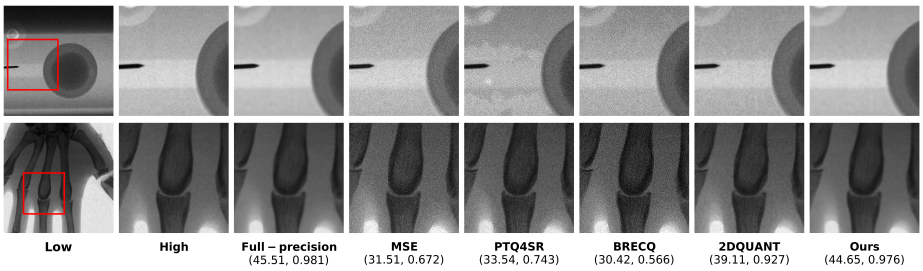

Fig. 4. Qualitative comparison on fluoroscopy (4-bit, SwinIR backbone). Visual results and corresponding PSNR and SSIM values for 4-bit PTQ.

Table 1 reports quantitative results (PSNR/SSIM). Across all architectures, conventional PTQ methods exhibit substantial fidelity loss under ultra-low precision, especially at 2-bit. In contrast, our method consistently maintains reconstruction fidelity across bit-widths, approaching full-precision performance at 4-bit. Compared to the strongest baseline (2DQuant), our method improves PSNR by up to 1.79 dB on CT and 5.54 dB on fluoroscopy. The improvements remain consistent across backbones, supporting architecture-agnostic robustness.

Qualitative comparisons in Figure 3 and Figure 4 further support these findings. Conventional PTQ methods often exhibit stronger quantization-induced noise and artifacts, whereas our approach alleviates these high-frequency distortions while maintaining overall reconstruction fidelity.

Deployment Efficiency under ONNX Runtime. We report ONNX-exported model sizes and compression ratios. As shown in Table 2, our method achieves compression up to 3.86× for INT8 and 7.40× for INT4 while preserving reconstruction fidelity.

3.3 Ablation Study

Effect of Dual-Stage Optimization. We evaluate the proposed dual-stage framework by comparing Stage 1 only with the full method (Stage 1 + Stage 2). As shown in Table 3, adding the wavelet-guided refinement consistently improves reconstruction fidelity across architectures and datasets, yielding up to 14.4 dB PSNR gain. Figure 5 further shows that this improvement is accompanied by

Table 3. Ablation study of frequency-aware refinement (2-bit). For PSNR and SSIM, values are reported as Stage 1/Full method, with Δ indicating the improvement.

Model	Mayo CT				Fluoroscopy			
	PSNR	SSIM	ΔPSNR	ΔSSIM	PSNR	SSIM	ΔPSNR	ΔSSIM
NAFNet	30.4/36.1	0.67/0.89	+5.7	+0.22	38.9/42.5	0.95/0.97	+3.6	+0.02
SwinIR	33.4/36.4	0.80/0.89	+3.0	+0.09	30.3/41.9	0.73/0.96	+11.6	+0.23
Restormer	34.6/37.3	0.85/0.91	+2.7	+0.06	28.0/42.4	0.92/0.97	+14.4	+0.05

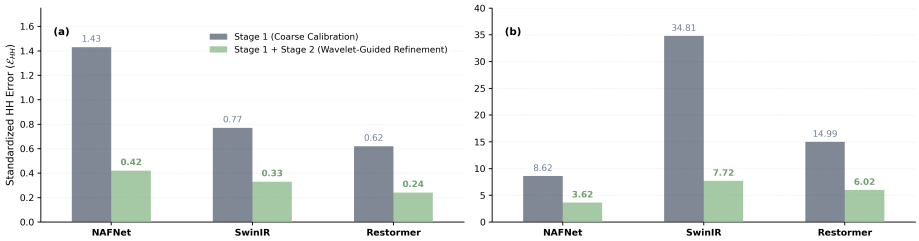

Fig. 5. Reduction of standardized HH subband error ($\mathcal{E}_{HH}$) after wavelet-guided refinement. Comparison between Stage 1 and the full method across backbones on (a) CT and (b) fluoroscopy datasets.

a substantial reduction in the standardized HH subband error $\mathcal{E}_{HH}$, indicating that the proposed method effectively mitigates high-frequency distortion.

Analysis of Empirical Design Choices. To examine the effect of empirical design choices, we evaluate calibration-set size and subband-weight variants under the 2-bit NAFNet setting. As shown in Table 4, the proposed refinement maintains strong performance across tested weight ratios, and increasing the calibration set beyond 10 triplets per patient provides limited additional gains. These results suggest that the method remains effective under moderate variations in calibration size and subband weighting.

4 Conclusion and Future Work

In this paper, we presented the first frequency-aware post-training quantization (PTQ) framework for medical image denoising. Through stationary wavelet transform (SWT) analysis, we showed that ultra-low-bit quantization induces frequency-selective distortion, with errors disproportionately concentrated in high-frequency subbands. Motivated by this observation, we introduced a dual-stage optimization strategy that explicitly suppresses subband-level distortion. The proposed method consistently preserves reconstruction fidelity at 2–4 bit precision across CNN, Transformer, and hybrid denoisers, while achieving substantial model-size reductions.

Future work will extend PTQ to medical restoration tasks beyond denoising and to emerging high-capacity denoisers such as diffusion-based models. We will also investigate reader- or task-level evaluation for clinically oriented assessment.

Table 4. Analysis of calibration-set size and subband weights. PSNR/SSIM are reported for 2-bit NAFNet. Triplets are per patient; Ratio is $\lambda_{LL} : \lambda_{HF} : \lambda_{HH}$.

(a) Calibration-set size

Triplets	CT	Fluoroscopy
5	35.98/0.884	41.92/0.970
10 (main)	36.08/0.891	42.50/0.973
20	36.12/0.899	42.51/0.973

(b) Subband weights

Ratio	CT	Fluoroscopy
1 : 1 : 1	36.11/0.888	42.21/0.972
2 : 3 : 5 (main)	36.08/0.891	42.50/0.973
1 : 3 : 6	36.09/0.891	43.48/0.970

Acknowledgments. This work was supported by grants from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (Nos. RS-2025-02215813, RS-2025-00520578, and RS-2025-02634603) and by an NRF grant funded by the Korean government (MSIT and MOE) (No. RS-2025-16063688).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brenner, D.J., Hall, E.J.: Computed tomography–an increasing source of radiation exposure. The New England journal of medicine **357**(22), 2277–2284 (2007)
2. McCollough, C.H., Primak, A.N., Braun, N., Kofler, J., Yu, L., Christner, J.: Strategies for reducing radiation dose in CT. Radiologic clinics of North America **47**(1), 27–40 (2009). https://doi.org/10.1016/j.rcl.2008.10.006
3. Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G.: Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. IEEE transactions on medical imaging **36**(12), 2524–2535 (2017). https://doi.org/10.1109/TMI.2017.2715284
4. Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., Yu, H.: CTformer: convolution-free Token2Token dilated vision transformer for low-dose CT denoising. Physics in Medicine & Biology **68**(6), 065012 (2023)
5. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5728–5739 (2022)
6. Yuan, J., Zhou, F., Guo, Z., Li, X., Yu, H.: HCformer: hybrid CNN-transformer for LDCT image denoising. Journal of Digital Imaging **36**(5), 2290–2305 (2023)
7. Chen, Z., Gao, Q., Zhang, Y., Shan, H.: ASCON: Anatomy-aware supervised contrastive learning framework for low-dose CT denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 355–365 (2023)
8. Zhang, J., Chao, H., Xu, X., Niu, C., Wang, G., Yan, P.: Task-oriented low-dose CT image denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 441–450 (2021)
9. AAPM Low Dose CT Grand Challenge, https://www.aapm.org/grandchallenge/lowdosect/, last accessed 2026/02/27
10. Jeon, S.Y., Wang, S., Wang, A.S., Gold, G.E., Choi, J.H.: Unsupervised training of a dynamic context-aware deep denoising framework for low-dose fluoroscopic imaging. IEEE Transactions on Instrumentation and Measurement (2025)

11. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European conference on computer vision, pp. 17–33 (2022)
12. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision Workshops, pp. 1833–1844 (2021)
13. Nagel, M., Amjad, R.A., Van Baalen, M., Louizos, C., Blankevoort, T.: Up or down? adaptive rounding for post-training quantization. In: International conference on machine learning, pp. 7197–7206 (2020)
14. Wang, H., Lu, J., Zhang, Y., Fu, Y.: Outlier-Aware Post-Training Quantization for Image Super-Resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16175–16184 (2025)
15. Tu, Z., Hu, J., Chen, H., Wang, Y.: Toward accurate post-training quantization for image super resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5856–5865 (2023)
16. Qin, H., Zhang, Y., Ding, Y., Liu, Y., Liu, X., Danelljan, M., Yu, F.: QuantSR: accurate low-bit quantization for efficient image super-resolution. Advances in neural information processing systems, 36, 56838–56848 (2023)
17. Liu, K., Qin, H., Guo, Y., Yuan, X., Kong, L., Chen, G., Zhang, Y.: 2DQuant: Low-bit post-training quantization for image super-resolution. Advances in Neural Information Processing Systems, 37, 71068–71084 (2024)
18. Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., Gu, S.: Brecq: Pushing the limit of post-training quantization by block reconstruction (2021). https://arxiv.org/abs/2102.05426
19. Zhang, R., Chung, A.C.: MedQ: Lossless ultra-low-bit neural network quantization for medical image segmentation. Medical Image Analysis **73**, 102200 (2021)
20. Shen, Y., Li, J., Shao, X., Inigo Romillo, B., Jindal, A., Dreizin, D., Unberath, M.: FastSAM3D: An Efficient Segment Anything Model for 3D Volumetric Medical Images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 542–552 (2024)
21. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2704–2713 (2018)
22. Nason, G.P., Silverman, B.W.: The stationary wavelet transform and some statistical applications. In: Wavelets and statistics, **103**, 281–299 (1995).
23. Pesquet, J.C., Krim, H., Carfantan, H.: Time-invariant orthonormal wavelet representations. IEEE transactions on signal processing, **44**(8), 1964–1970 (1996)
24. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the Spectral Bias of Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning, **97**, 5301–5310 (2019)
25. Burgess, A.E., Jacobson, F.L., Judy, P.F.: Detection of lesions in mammographic structure. In: Medical Imaging 1999: Image Perception and Performance, **3663**, 304–315 (1999)
26. Chen, L., Abbey, C.K., Boone, J.M.: Association between power law coefficients of the anatomical noise power spectrum and lesion detectability in breast imaging modalities. Physics in medicine and biology, **58**(6), 1663–1681 (2013)
27. Kim, W., Lee, J., Kang, M., Kim, J.S., Choi, J.H.: Wavelet subband-specific learning for low-dose computed tomography denoising. PLoS ONE, **17**(9): e0274308 (2022). https://doi.org/10.1371/journal.pone.0274308

28. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or -1 (2016). https://arxiv.org/abs/1602.02830