FREQUENCY MATTERS: FREQUENCY-AWARE DECOUPLING AND HARMONIZED ENHANCEMENT FOR HIGH-FIDELITY MEDICAL IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image super-resolution (SISR) in medical imaging is hindered by the low-frequency bias of deep learning and the distinct semantic roles of mid- and high-frequency components, which uniform enhancement strategies fail to address. We propose the Frequency-Aware Representation Enhancer (FARE), a plug-and-play module that explicitly decouples and enhances these frequency bands. FARE comprises a High-Frequency Semantic Amplifier (HFSA), which combines directional wavelet decomposition with gradient-informed gating, and an Adaptive Mid-Frequency Decoupler (AMFD), which applies patch-level STFT with a learnable Gaussian band-pass mask. Their outputs are fused via cross-frequency negotiated gating. Extensive evaluations on multiple medical image benchmarks demonstrate consistent improvements over state-of-the-art methods.

1 INTRODUCTION

Single-image super-resolution (SISR) has been widely applied across various domains, with medical imaging analysis tasks being a key application. In this context, SISR enhances the spatial resolution of diagnostic scans, facilitating the visualization of fine anatomical structures and supporting clinical decision-making. Early methods relied on interpolation (Keys, 2003), Markov random fields (He & Siu, 2011), and sparse dictionary learning (Yang et al., 2010). The advent of deep learning (DL) has significantly advanced the field by enabling the modeling of complex mappings from low-resolution (LR) to high-resolution (HR) images, resulting in substantial improvements in reconstruction quality (Dong et al., 2015; Dai et al., 2019; Kim et al., 2016; Lim et al., 2017; Zhang et al., 2018a). These approaches often integrate residual learning and attention mechanisms to refine texture and perceptual fidelity. Recently, transformer-based architectures leveraging self-attention have further improved performance by effectively capturing long-range dependencies (Liang et al., 2021; Li et al., 2023; Chen et al., 2023).

Despite the notable improvements in clarity achieved by DL-based super-resolution (SR) models, SISR continues to face challenges in reconstructing fine structural details, a limitation of particular concern in medical imaging, where the accurate recovery of intricate anatomical features is crucial. A key contributing factor is the low-frequency bias inherent in DL-based models (Molina et al., 2024; Lin et al., 2022; Fridovich-Keil et al., 2022), which often leads to the underrepresentation of essential structures embedded in mid- and high-frequency components, such as complex textures and subtle anatomical boundaries (Azad et al., 2021). This bias is further amplified in medical images due to their inherently polarized frequency distributions (Qian et al., 2025), which arise from the coexistence of large, homogeneous regions and densely packed anatomical details. Recent studies have explored frequency-domain modeling using wavelet and Fourier-based techniques (Huang et al., 2017; Jeevan et al., 2024; Fuoli et al., 2021). However, growing evidence indicates that mid- and high-frequency components in medical images convey distinct semantic information (Azad et al., 2021; Bo et al., 2025; Huang et al., 2021), rendering unified processing strategies inadequate. Consequently, methods relying solely on high-frequency subbands or global frequency loss functions often fail to capture these semantic distinctions. Moreover, as shown in Figure 1, the wide distribution ranges of texture descriptors, such as Second Moment, GLCM Contrast, Energy, and Entropy,

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.

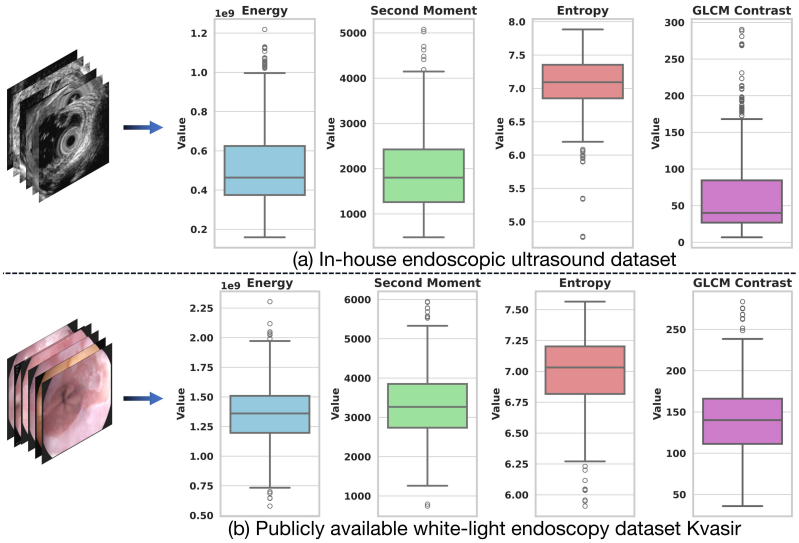


Figure 1: Statistical distributions of representative texture descriptors. Second Moment and GLCM Contrast capture texture intensity and local gray-level variation. Energy and Entropy reflect high-frequency detail and distributional uncertainty, with Entropy serving as a multi-frequency complexity measure. Their broad ranges indicate substantial texture variability across the dataset.

highlight significant variability in structural complexity across images within the same dataset. This motivates the need for a dynamic partitioning mechanism that adaptively identifies mid-frequency intervals and decouples frequency components on a per-image basis.

In this study, we propose the Frequency-Aware Representation Enhancer (FARE), a novel plug-and-play module for medical image super-resolution that explicitly decouples mid- and high-frequency components and adaptively enhances them through harmonized interaction. FARE consists of the High-Frequency Semantic Amplifier (HFSA) and the Adaptive Mid-Frequency Decoupler (AMFD), which respectively enhance high-frequency structures and dynamically refine mid-frequency information. Specifically, HFSA employs DTCWT to exploit its compact support, directional selectivity, and spatial localization to capture abrupt variations and anatomical boundaries. A gradient-based prior further guides the model toward informative high-frequency cues while suppressing spurious fluctuations. However, wavelet subbands are generated by finite-support filters rather than strict rectangular partitions in the Fourier domain, resulting in transition regions and partial overlap between adjacent bands. This makes wavelet decomposition less suitable for explicitly isolating an adjustable mid-frequency interval. Moreover, its strong spatial-localization bias limits the explicit modeling of long-range interactions required by coherent tissue textures and gradual anatomical patterns. Therefore, AMFD adopts patch-wise STFT with a learnable Gaussian band-pass mask, which adaptively controls the center and bandwidth of the selected frequency interval. The Fourier representation provides broader contextual spectral modeling within each window, while Overlap-Add preserves continuity across neighboring patches. Finally, HFSA and AMFD are integrated through an interleaved negotiation-fusion mechanism with gated integration and residual connections. This design coordinates complementary frequency-specific representations, mitigates cross-band interference, and prevents either branch from dominating. As a result, FARE enables efficient and harmonized frequency-aware representation learning in a modular form. Our contributions are listed as follows:

- We propose FARE, a plug-and-play frequency-aware enhancement module for medical image super-resolution, featuring a novel negotiation-fusion mechanism.
- A newly developed HFSA submodule integrates wavelet-derived high-frequency subbands with gradient features to selectively enhance high-frequency structural details.
- We introduce the AMFD submodule, which dynamically generates band-pass masks to enable adaptive decoupling and modeling of mid-frequency features.
- We conduct extensive experiments across multiple medical image SR tasks to validate the effectiveness of FARE and the complementary roles of each submodule.

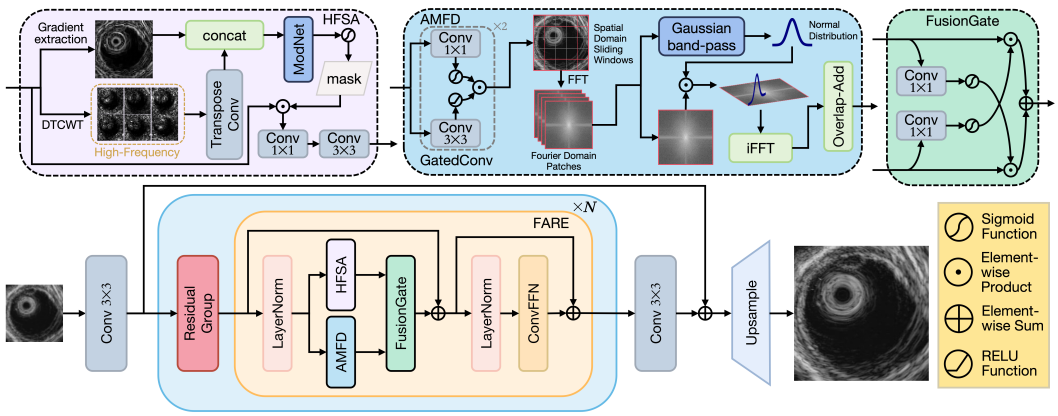


Figure 2: Overview of the proposed architecture. The backbone adopts a residual group-based super-resolution framework. The proposed FARE is a plug-and-play module, which comprises two key components: the HFSA and AMFD.

2 METHODOLOGY

2.1 OVERALL ARCHITECTURE

Most modern SR networks employ a modular architecture composed of multiple Residual Groups (RGs), each containing convolutional or attention-based blocks that progressively refine hierarchical spatial features. Leveraging this paradigm, we propose the FARE, a generic plug-in module designed for seamless integration after each RG. Specifically, given an input low-resolution (LR) image $I_L \in \mathbb{R}^{B\times C\times H\times W}$, shallow features are first extracted and subsequently passed through a backbone network comprising N stacked RGs. After each RG, the intermediate feature X is processed by the FARE to enhance frequency-aware representations before being forwarded to the next stage. This iterative integration enables progressive accumulation of frequency-sensitive refinements throughout the network depth. Accordingly, FARE is architecture-agnostic, wherein the fully feed-forward and differentiable structure requires no modification to the backbone architecture, operating independently of the host network's internal design. This design allows effortless integration into a wide range of existing or future SR frameworks. Additionally, FARE preserves feature dimensionality and continuity, ensuring compatibility with residual connections and reconstruction heads. This property makes it a practical and scalable solution for frequency-aware enhancement in contemporary SR systems.

As illustrated in Figure 2, the FARE includes two complementary modules: the HFSA and AMFD, each targeting distinct frequency characteristics and associated modeling challenges. Specifically, HFSA employs wavelet-based spatial-frequency localization to selectively enhance sharp structural transitions, thereby emphasizing salient high-frequency information. In contrast, AMFD conducts dynamic mid-frequency decomposition through short-time spectral analysis, adaptively isolating and refining contextual textures. Subsequently, the outputs of these two branches are integrated via a cross-frequency negotiated gating mechanism, facilitating mutual semantic calibration between frequency domains. The resulting enhanced representation F_{FARE} is then propagated to the subsequent RG or the final reconstruction head, enabling fine-grained detail restoration within a frequency-aware modeling framework.

2.2 HIGH-FREQUENCY SEMANTIC AMPLIFIER

High-frequency components encode critical structural information, which is especially indispensable for precise clinical evaluation. However, the inherent low-frequency bias of conventional neural architectures impedes the effective use of these critical high-frequency components. It is essential to adopt representations that are both sensitive to high-frequency variations and spatially localized. Wavelet transforms offer a principled solution with their compactly supported basis functions, enabling precise localization of abrupt structural changes while preserving directional information. Building on this property, we introduce the HFSA module to explicitly enhance high-frequency responses through localized and directional wavelet analysis. Specifically, we employ the Dual-Tree

Complex Wavelet Transform (DTCWT) to decompose the input feature map X into low- and high-frequency subbands:

$$y_l, \{y_h^k\}_{k=1}^6 = \mathrm{DTCWT}(X), \tag{1}$$

where y_l is the low-frequency subband and $\{y_h^k\}_{k=1}^6$ are the high-frequency components oriented at $\pm 15°, \pm 45°, and \pm 75°$. These directional subbands are particularly well-suited for modeling local structural discontinuities, as the wavelet bases facilitate fine-grained spatial localization of sharp intensity transitions.

The magnitude maps of the high-frequency subbands are upsampled using transposed convolution for subsequent feature integration. However, high-frequency subbands may also contain unstructured noise, which can degrade the quality of feature representations. Moreover, the use of fixed, explicit decomposition limits the model's ability to capture semantic context adaptively. To address these challenges, we incorporate a structure-aware prior derived from the image's gradient map, which encodes local intensity variations indicative of anatomical boundaries. This gradient-based prior serves as a spatial cue to help the model distinguish meaningful structural transitions from spurious high-frequency noise fluctuations, calculated as follows:

$$E = \sqrt{(\partial_x X)^2 + (\partial_y X)^2 + \epsilon}, \quad \hat{E} = \frac{E}{\max(E)}, \tag{2}$$

where ∂_x and ∂_y denote the spatial derivatives, and $\hat{E}$ represents the normalized gradient map.

The structural prior $\hat{E}$ is concatenated with the high-frequency subbands upsampled via transposed convolution, and then passed through a shallow convolutional gating network to produce a spatial attention mask. By integrating gradient-based priors with the directional selectivity of wavelet features, the network is guided to modulate the high-frequency responses of the input features more effectively, emphasizing semantically relevant structures while suppressing noise and redundancy. Finally, the features are fused through a 1×1 convolution followed by a 3×3 convolution to generate the output of the HFSA module. This hybrid strategy enhances both the adaptability and robustness of the model in capturing anatomical details, while effectively suppressing irrelevant or noise-induced activations.

2.3 Adaptive Mid-Frequency Decoupler

Semantic information manifests differently across frequency bands, particularly encoding essential structural and contextual features in medical imaging (Azad et al., 2021; Bo et al., 2025; Huang et al., 2021). Moreover, as illustrated in Figure 1, texture-related attributes exhibit substantial variability even within the same dataset, indicating that fixed mid-frequency thresholds are inadequate for defining universally applicable frequency bands. Therefore, we propose AMFD to decouple mid-frequency information from input features by learning an adaptive band-pass filter.

Mainly, the AMFD module begins with two gated convolutional layers designed to suppress redundant or semantically irrelevant regions, thereby reducing the computational overhead of subsequent frequency-domain operations. These spatial gates serve as early-stage filters that constrain spectral analysis to the most informative areas. Subsequently, overlapping patches are extracted using a sliding-window implemented via the Unfold operation:

$$P = \mathrm{Unfold}_{w,h}(s_2) \in \mathbb{R}^{B \times Cw^2 \times L}, \tag{3}$$

where w denotes the window size and L is the total number of extracted patches. Each patch $p_j \in \mathbb{R}^{C \times w \times w}$ is transformed into the frequency domain via the Fourier Transform, following the principle of STFT:

$$X_j = \mathcal{F}(p_j), \quad M_j = |X_j|, \tag{4}$$

where $|\cdot|$ denotes the magnitude spectrum. A Gaussian-shaped mask is applied in the frequency domain to dynamically isolate mid-frequency components. To enhance numerical stability during optimization, we adopt an affine reparameterization of the standard Gaussian expression. Specifically, we transform M_j into a normalized form using learnable parameters μ and σ, where μ denotes the center and σ the bandwidth of the frequency band. The reparameterized formulation is expressed as:

$$\begin{cases} y = \dfrac{M_j - \mu}{\sigma} = \dfrac{1}{\sigma} M_j - \dfrac{\mu}{\sigma} = wM_j + b, \\ M_j^{\text{mask}} = e^{\left(-\frac{(M_j-\mu)^2}{2\sigma^2}\right)} = e^{\left(-\frac{1}{2}(wM_j+b)^2\right)}. \end{cases} \tag{5}$$

Specifically, the parameters w and b are dynamically predicted from the input patch via 1×1 convolutions. This formulation enables the transformation of spectral representations into a learnable and normalized space, facilitating adaptive and precise control over the band-pass window.

Finally, the masked spectra are transformed back to the spatial domain via the inverse Fourier Transform and aggregated through the Overlap-Add method:

$$\hat{s}_2 = \frac{\text{Fold}(\mathcal{F}^{-1}(X_j \odot M_j^{\text{mask}}))}{\text{Fold}(\text{Unfold}(\mathbf{1}))}. \tag{6}$$

This reconstruction strategy preserves spatial consistency while effectively enhancing mid-frequency representations during the reconstruction process. Coupled with the learnable spectral masking, the AMFD module offers a practical and adaptive approach for decoupling and refining mid-frequency components.

2.4 Cross-Frequency Negotiated Gating and Fusion

Rather than treating the outputs of HFSA and AMFD independently, we introduce a cross-frequency negotiated gating mechanism to enable mutual guidance and collaborative integration. This design mitigates the risk of dominance by a single frequency component and promotes a more balanced, semantically enriched representation. Specifically, each branch generates an attention weight for the other, conditioned on the features extracted by its counterpart:

$$g_m = \sigma(W_m * F_{\text{HFSA}}), \quad g_h = \sigma(W_e * F_{\text{AMFD}}), \tag{7}$$

where W_m and W_e denote 1×1 convolution kernels. These attention weights are subsequently employed to modulate and fuse the frequency-specific representations in a mutually guided manner:

$$F_{\text{FARE}} = F_{\text{AMFD}} \odot g_m + F_{\text{HFSA}} \odot g_h. \tag{8}$$

This bidirectional attention framework enables dynamic negotiation between mid- and high-frequency cues, resulting in a harmonized and context-aware feature representation.

Finally, the fused representation F_{FARE} is passed through a convolutional feed-forward network to enhance nonlinearity and representational capacity before the subsequent RG or final image reconstruction. As a self-contained and modular unit, FARE is agnostic to specific backbone architectures, which can be seamlessly integrated into a wide range of super-resolution networks as a plug-and-play component for frequency-aware semantic enhancement.

3 Experiment

3.1 Experiment Setting

Datasets. To expand the applicability across different modalities, we evaluate our method on multiple datasets. We adopt two widely used medical image datasets for benchmarking. The HC18 dataset (van den Heuvel et al., 2018) is a fetal head ultrasound dataset. It consists of clinical ultrasound images from various sources and contains a wide variety of anatomical appearances and noise characteristics. The Kvasir dataset (Pogorelov et al., 2017) is a gastrointestinal endoscopy dataset. It includes images covering multiple lesion types, which are essential for evaluating perceptual detail restoration under complex textural scenarios.

We also evaluate our method on an in-house endoscopic ultrasound dataset collected from hospital. The data was acquired using Olympus Endoscopic Ultrasound systems between 2011-05 and 2020-10. The study was approved by the institutional ethics committee. To ensure robust evaluation, the dataset is partitioned strictly by patient identity, preventing data of the same individual from appearing in both training and testing sets. This protocol guarantees unbiased performance assessment across different patients and anatomical variations.

Table 1: Quantitative results for the **blind** super-resolution task on three datasets. The better results are highlighted in **bold**.

Datasets	In-house			Kvasir			HC18		
Metrics	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RCAN (Zhang et al., 2018a)	27.66	0.7052	0.1559	30.48	0.8542	0.1287	27.03	0.7665	0.2018
RCAN+FARE	**27.98**	**0.7118**	**0.1525**	**31.07**	**0.8548**	**0.1222**	**27.33**	**0.7671**	**0.1801**
HAN (Niu et al., 2020)	26.63	0.7048	0.1390	31.47	0.8515	0.1225	27.21	0.7743	0.1851
HAN+FARE	**27.03**	**0.7057**	**0.1327**	**31.74**	**0.8516**	**0.1196**	**27.70**	**0.7835**	**0.1742**
SwinIR (Liang et al., 2021)	29.33	0.7490	0.1386	34.46	0.8753	0.1112	26.31	0.7209	0.1629
SwinIR+FARE	**29.44**	**0.7516**	**0.1353**	**34.55**	**0.8755**	**0.1102**	**26.44**	**0.7330**	**0.1578**
ART (Zhang et al., 2022)	29.36	0.7474	0.1432	34.70	0.8768	0.1101	28.20	0.7831	0.1354
ART+FARE	**29.52**	**0.7540**	**0.1336**	**34.82**	**0.8770**	**0.1096**	**28.21**	**0.7834**	**0.1353**
ATD (Zhang et al., 2024)	29.40	0.7484	0.1412	34.61	0.8764	0.1108	28.52	0.7959	0.1744
ATD+FARE	**29.54**	**0.7541**	**0.1359**	**34.81**	**0.8774**	**0.1105**	**28.76**	**0.8012**	**0.1669**
PFT (Long et al., 2025)	29.46	0.7517	0.1380	34.53	0.8762	0.1102	27.94	0.7697	0.1116
PFT+FARE	**29.49**	**0.7527**	**0.1371**	**34.55**	**0.8764**	**0.1101**	**28.03**	**0.7716**	**0.1105**

For MRI evaluation, we use two multi-contrast datasets, BraTS2018 and IXI. BraTS2018 comprises multimodal MRI scans, and IXI contains multi-contrast MRI volumes. In our experiments, we use T1 as the reference to reconstruct the target T2 modality for BraTS2018, whereas PD serves as the reference to reconstruct T2 for IXI. Both datasets are split into training, validation, and test sets in a 7:1:2 ratio. For each volume, only the central 100 axial slices are used.

Degradation Settings. We adopt both blind and non-blind super-resolution settings to evaluate model performance under different degradation conditions. Blind super-resolution aims to restore HR images from LR inputs with unknown and complex degradations, while non-blind super-resolution assumes a known and fixed degradation process. For the blind super-resolution task, we adopt the BSRGAN degradation method to generate LR images, simulating the quality degradation caused by imaging equipment, noise, and other uncontrollable factors in real clinical scenarios. For the non-blind super-resolution task, we use standard bicubic interpolation downsampling as the degradation process, focusing on evaluating the model's restoration ability under ideal downsampling conditions. All downsampling operations are performed with a $\times 4$ scale factor. For MRI, the downsampling is performed in the k-space domain, which simulates a more realistic MRI reconstruction scenario. We use a $\times 4$ downsampling factor, where the k-space data is subsampled and then used for reconstruction, further ensuring that the metrics are aligned with MRI-specific restoration challenges.

Evaluation and Implementation. We adopt six representative state-of-the-art super-resolution models as baseline methods, including RCAN (Zhang et al., 2018a), HAN (Niu et al., 2020), SwinIR (Liang et al., 2021), ART (Zhang et al., 2022), ATD (Zhang et al., 2024), and PFT (Long et al., 2025). The proposed FARE module is integrated into each of these architectures in a plug-and-play manner, allowing for direct performance comparison with their original backbone. The evaluation metrics include PSNR and SSIM, which are calculated in the luminance channel after the image is converted to YCbCr space. LPIPS is also introduced to measure perceptual similarity and is computed directly on the RGB channels. We use the Adam optimizer with an initial learning rate of 2×10^{-4}, which is decayed by a factor of 0.5 at 150k, 200k, 225k, and 240k iterations, respectively. The total training iterations is 250k. The loss function is the pixel-wise L1 loss. All experiments are conducted on NVIDIA RTX 4090 GPU.

3.2 Comparison With Baseline Methods

Blind Super-Resolution Results. As shown in Table 1, the proposed FARE module consistently improves performance across all baseline models and datasets under the blind super-resolution setting. Notably, both PSNR and SSIM increase while LPIPS significantly decreases, indicating that FARE not only enhances pixel-level fidelity but also improves perceptual quality. Taking the ATD model (Zhang et al., 2024) as an example. On the in-house dataset, ATD+FARE achieves a PSNR

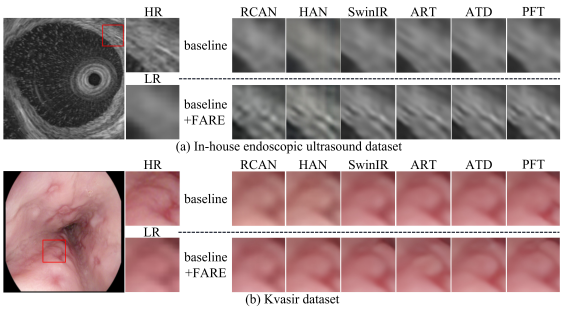


Figure 3: Visual comparison of the **blind** super-resolution task.

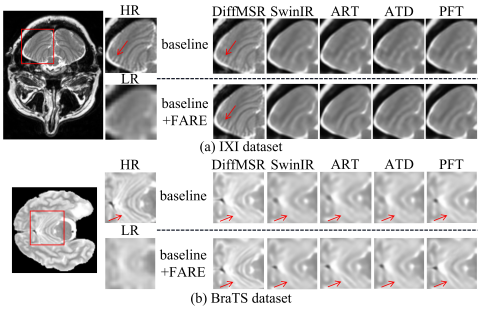


Figure 4: Visual comparison of the **MRI** super-resolution task.

improvement of 0.14 dB, SSIM gain of 0.0057, and LPIPS reduction of 0.0053. On the HC18 fetal ultrasound dataset, FARE boosts SSIM by 0.0053 and reduces LPIPS by 0.0075, highlighting its ability to recover semantically meaningful anatomical structures with higher perceptual similarity to ground truth. Even on the Kvasir dataset, which exhibits relatively low structural complexity, FARE brings an additional 0.20 dB in PSNR, demonstrating its robustness across varying anatomical and texture conditions. Moreover, models like RCAN and HAN also benefit notably. RCAN+FARE gains 0.59 dB in PSNR on Kvasir, while HAN+FARE delivers the largest LPIPS reduction of 0.0109 on the HC18 dataset. These improvements are achieved without altering the original network architecture, underscoring the plug-and-play design of FARE. Consistent gains across all backbones and datasets validate the broad applicability and effectiveness of the proposed frequency-aware enhancement strategy.

Additionally, as shown in Figure 3, we visualize representative results from the blind super-resolution experiments on both the in-house and Kvasir datasets. These comparisons highlight the effectiveness of the proposed FARE module in enhancing visual quality across diverse anatomical contexts. In (a), baseline models produce oversmoothed textures that fail to preserve fine structural variations. In contrast, models enhanced with FARE counterparts restore sharper and more faithful tissue textures. In (b), baseline reconstructions exhibit blurry vascular boundaries, while models with FARE integration successfully recover more distinct vessel contours. These results indicate that FARE can effectively guide the network to focus on semantically meaningful frequency components, thereby enhancing perceptual detail and yielding more precise reconstructions.

To further assess the frequency-domain effectiveness of our method, we evaluate reconstructed images using four spectral similarity metrics. Residual energy (RE) quantifies the overall deviation between the spectra of reconstructed and ground-truth images, computed via their ℓ_2 distance. Spectral correlation coefficient (SCC) measures the linear correlation between these two spectra, reflecting structural similarity in the frequency domain. KL divergence (KLD) evaluates the asymmetric discrepancy between these two spectra by interpreting the latter as the target distribution. Mid- and high-frequency energy ratio (MHER) evaluates the proportion of spectral energy located beyond the low-frequency core, thereby reflecting the reconstruction quality of mid- and high-frequency components. Detailed formulas are provided in the **Appendix**. As shown in Figure 5, our method achieves consistent improvements across all four metrics, indicating enhanced frequency-domain alignment and structural fidelity.

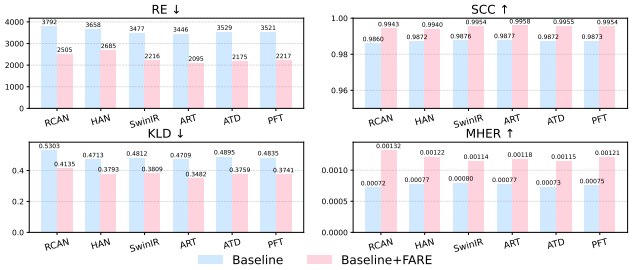


Figure 5: Quantitative results for the frequency-domain evaluation under the **blind** super-resolution setting. All four metrics are computed by averaging over the entire reconstructed results on the in-house dataset.

Non-Blind Super-Resolution Results. We also evaluate the generality of the proposed module on the non-blind super-resolution task. By integrating FARE into the aforementioned baseline networks, the quantitative results are shown in Table 2. It can be observed that after incorporating FARE into all networks, PSNR and SSIM are consistently improved, and LPIPS further decreases.

Table 2: Quantitative results for the **non-blind** super-resolution task on three datasets. The better results are highlighted in **bold**.

Datasets	In-house			Kvasir			HC18		
Metrics	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RCAN (Zhang et al., 2018a)	29.03	0.8250	0.0768	32.84	0.8887	0.0889	30.19	0.8719	0.1393
RCAN+FARE	**31.11**	**0.8292**	**0.0651**	**35.22**	**0.8950**	**0.0588**	**31.27**	**0.8767**	**0.1351**
HAN (Niu et al., 2020)	30.58	0.8433	0.0562	35.59	0.8978	0.0562	30.82	0.8671	0.1011
HAN+FARE	**31.07**	**0.8465**	**0.0553**	**35.62**	**0.8979**	**0.0559**	**32.59**	**0.8924**	**0.0985**
SwinIR (Liang et al., 2021)	33.90	0.8736	0.0552	39.60	0.9365	0.0388	33.48	0.8951	0.0678
SwinIR+FARE	**34.02**	**0.8762**	**0.0551**	**39.62**	**0.9372**	**0.0364**	**33.60**	**0.8963**	**0.0666**
ART (Zhang et al., 2022)	34.50	0.8845	0.0583	39.26	0.9331	0.0414	**33.41**	0.8906	0.0566
ART+FARE	**34.51**	**0.8853**	**0.0570**	**39.58**	**0.9367**	**0.0373**	**33.41**	**0.8912**	**0.0563**
ATD (Zhang et al., 2024)	34.50	0.8849	0.0572	40.04	0.9406	0.0352	33.82	0.9037	0.0783
ATD+FARE	**34.53**	**0.8857**	**0.0563**	**40.13**	**0.9417**	**0.0335**	**33.93**	**0.9044**	**0.0778**
PFT (Long et al., 2025)	34.43	0.8827	0.0585	38.89	0.9295	0.0421	33.37	0.8898	0.0520
PFT+FARE	**34.44**	**0.8834**	**0.0574**	**39.43**	**0.9355**	**0.0379**	**33.51**	**0.8918**	**0.0505**

Table 3: Quantitative results of MRI super-resolution on BraTS and IXI. The better results are highlighted in **bold**.

Datasets	BraTS			IXI		
Metrics	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SwinIR (Liang et al., 2021)	31.28	0.9561	0.0690	25.39	0.8427	0.1791
SwinIR+FARE	**31.81**	**0.9580**	**0.0587**	**25.68**	**0.8503**	**0.1709**
ART (Zhang et al., 2022)	31.65	0.9567	0.0590	25.73	0.8520	0.1697
ART+FARE	**32.07**	**0.9594**	**0.0512**	**26.13**	**0.8624**	**0.1602**
ATD (Zhang et al., 2024)	30.91	0.9517	0.0662	24.94	0.8297	0.1900
ATD+FARE	**31.66**	**0.9569**	**0.0539**	**25.56**	**0.8476**	**0.1714**
PFT (Long et al., 2025)	32.25	0.9605	0.0521	25.12	0.8356	0.1914
PFT+FARE	**32.62**	**0.9637**	**0.0454**	**25.55**	**0.8477**	**0.1753**
DiffMSR (Li et al., 2024)	35.78	0.9751	0.0322	33.44	0.9544	0.0773
DiffMSR+FARE	**36.12**	**0.9780**	**0.0258**	**33.82**	**0.9553**	**0.0745**

This demonstrates that the proposed frequency-domain enhancement module is not only effective in complex degradation scenarios but also significantly improves detail reconstruction and perceptual quality under standard ideal downsampling.

MRI Super-Resolution Results. As summarized in Table 3, FARE consistently improves PSNR and SSIM and reduces LPIPS for every backbone across both datasets; for example, SwinIR achieves PSNR gains of 0.53dB on BraTS and 0.29dB on IXI with corresponding SSIM increases and LPIPS drops. We also integrate it into the multi-contrast method DiffMSR (Li et al., 2024) by placing the module on intra-network feature blocks while keeping the original cross-contrast fusion unchanged. DiffMSR achieves PSNR gains of 0.34 dB on BraTS and 0.38 dB on IXI. Visual comparisons in Figure 4 demonstrate the noticeable visual improvements achieved by our method. These results demonstrate that FARE, originally designed for SISR, can effectively be transferred to multi-contrast settings, enhancing MRI reconstruction quality, and confirm that frequency-aware decoupling and harmonized enhancement complement cross-contrast guidance, and that the module can migrate as a plug-and-play enhancer to multi-contrast super-resolution task.

3.3 Ablation Study

To validate the independent effectiveness and complementarity of two sub-modules, the HFSA and AMFD, in frequency-domain enhancement, we conduct a detailed ablation study using SwinIR and ATD as the baselines under the blind super-resolution setting on the medical datasets.

Table 4: Ablation study for the **blind** super-resolution task. We report (i) module ablations of FARE on SwinIR and ATD, (ii) structure-aware prior ablation by setting gradient guidance to all ones (grad=1), and (iii) replaceability tests by removing FARE and instead adopting common alternatives. The best results are highlighted in **bold**.

Backbone	Variant	In-house			Kvasir		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SwinIR	Baseline (w/o FARE)	29.33	0.7490	0.1386	34.46	0.8753	0.1112
	+HFSA only	29.35	0.7493	0.1365	34.53	**0.8755**	0.1112
	+HFSA (grad=1)	29.33	0.7491	0.1376	34.50	0.8754	0.1113
	+AMFD only	29.38	0.7497	0.1361	34.51	0.8754	0.1108
	Replace w/ Fourier loss (Fuoli et al., 2021)	29.39	0.7492	0.1383	34.51	0.8753	0.1109
	Replace w/ ESA (Liu et al., 2020)	29.34	0.7491	0.1387	34.49	0.8753	0.1113
	Replace w/ CoordAtt (Hou et al., 2021)	29.35	0.7491	0.1372	34.52	0.8751	0.1112
	+FARE	**29.44**	**0.7516**	**0.1353**	**34.55**	**0.8755**	**0.1102**
ATD	Baseline (w/o FARE)	29.40	0.7484	0.1412	34.61	0.8764	0.1108
	+HFSA only	29.42	0.7492	0.1402	34.67	0.8770	0.1108
	+HFSA (grad=1)	29.42	0.7490	0.1409	34.65	0.8767	0.1109
	+AMFD only	29.44	0.7502	0.1392	34.77	**0.8774**	0.1108
	Replace w/ Fourier loss (Fuoli et al., 2021)	29.43	0.7485	0.1410	34.70	0.8769	0.1110
	Replace w/ ESA (Liu et al., 2020)	29.41	0.7484	0.1409	34.66	0.8763	0.1109
	Replace w/ CoordAtt (Hou et al., 2021)	29.44	0.7491	0.1401	34.65	0.8771	0.1107
	+FARE	**29.54**	**0.7541**	**0.1359**	**34.81**	**0.8774**	**0.1105**

The ablation results are presented in Table 4. It can be observed that both the HFSA-only and AMFD-only configurations outperform the original baseline, demonstrating that both modules achieve design objectives and enhance super-resolution performance of networks. (Extended ablations on additional backbones are provided in supplementary.)

More importantly, the full integration of FARE leads to further gains in PSNR and SSIM along with reductions in LPIPS. This performance boost stems from the negotiated fusion mechanism, which enables dynamic and frequency-aware interaction between mid- and high-frequency pathways. Specifically, through the cross-frequency gating structure, each branch generates attention weights guided by the complementary sub-module, allowing HFSA to focus on refining semantically meaningful high-frequency structures while AMFD adaptively adjusts mid-frequency enhancement conditioned on high-frequency cues. This bidirectional feedback process facilitates mutual refinement and avoids redundancy or dominance of a single frequency band, yielding a harmonized representation that balances perceptual fidelity and detail accuracy. These results strongly support the effectiveness of our proposed frequency-aware decoupling and cross-modulation strategy.

We further ablate the structure-aware prior in HFSA by setting the gradient-guidance tensor to all ones (grad=1, i.e., removing the gradient prior). This consistently degrades relative to HFSA-only, with LPIPS worsening most notably, confirming that the gradient-guided structural prior is vital.

To examine whether the gains can be replicated by simpler alternatives, we remove FARE and instead adopt representative frequency-domain supervision (Fourier loss) or lightweight attention modules (ESA, CoordAtt). As shown in Table 4, all such variants underperform the +FARE results, suggesting that our MF–HF semantic decoupling and negotiated fusion design are more effective.

4 Conclusion

We propose FARE, a plug-and-play frequency-aware enhancement module for medical image super-resolution, designed to address the structural ambiguity arising from low-frequency bias and frequency-specific semantic divergence. FARE enables targeted decoupling and modeling of frequency-specific representations, which significantly improves reconstruction fidelity and perceptual quality. Through extensive experiments that simulate ideal and clinically realistic settings, our method demonstrates consistent improvements across diverse datasets. Moreover, FARE transfers effectively from SISR to multi-contrast super-resolution, delivering consistent gains. As an architecture-agnostic design, FARE presents a promising direction for frequency-aware modeling in both low- and high-level vision tasks, with potential for broader utility.

AI USE STATEMENT

Generative AI tools, specifically Large Language Models (LLMs), were used to improve the clarity and linguistic quality of the manuscript and to assist with code implementation. No LLMs were used for research idea generation or experimental design. All AI-assisted code was carefully reviewed and tested by the authors to ensure correctness, and all AI-assisted text was manually checked and revised. The authors take full responsibility for the final manuscript and experimental results.

ETHICS STATEMENT

The in-house study received approval from the institutional ethics committee under IRB No. 2019-SR-448 and was conducted in accordance with the principles of the Declaration of Helsinki. The dataset was partitioned strictly by patient identity so that images from the same patient did not appear in both the training and test sets.

REPRODUCIBILITY STATEMENT

The backbone models and checkpoints used in this work are publicly available, and the experimental benchmarks are standardized and publicly accessible. We will release the implementation code in a public GitHub repository. Additional experimental results are provided in the appendix to further facilitate reproduction and verification of our findings.

REFERENCES

Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, Kagan Incetan, Guliz Irem Gokceler, Muhammed Ali Simsek, Kivanc Ararat, Richard J Chen, Nicholas J Durr, Faisal Mahmood, et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Transactions on Medical Imaging*, 39(12):4297–4309, 2020.

Reza Azad, Afshin Bozorgpour, Maryam Asadi-Aghbolaghi, Dorit Merhof, and Sergio Escalera. Deep frequency re-calibration u-net for medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3274–3283, 2021.

Yuntian Bo, Yazhou Zhu, Lunbo Li, and Haofeng Zhang. Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1889–1897, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22367–22377, 2023.

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11065–11074, 2019.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pp. 184–199. Springer, 2014.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Peng Du, Hui Li, Han Xu, Paul Barom Jeon, Dongwook Lee, Daehyun Ji, Ran Yang, and Feng Zhu. Diffusion transformer meets multi-level wavelet spectrum for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19700–19710, 2025.

Sara Fridovich-Keil, Raphael Gontijo Lopes, and Rebecca Roelofs. Spectral bias in practice: The role of function frequency in generalization. *Advances in Neural Information Processing Systems*, 35:7368–7382, 2022.

Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2360–2369, 2021.

Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR 2011*, pp. 449–456. IEEE, 2011.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13713–13722, 2021.

Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697, 2017.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6891–6902, 2021.

Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5884–5892, 2024.

Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 2003.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Cansu Korkmaz, A Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5926–5936, 2024.

Dongheon Lee, Seokju Yun, and Youngmin Ro. Emulating self-attention with convolution for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24467–24477, 2025.

Marcelo Lerendegui, Kai Riemer, Georgios Papageorgiou, Bingxue Wang, Lachlan Arthur, Arthur Chavignon, Tao Zhang, Olivier Couture, Pingtong Huang, Md Ashikuzzaman, et al. Ultra-sr challenge: Assessment of ultrasound localization and tracking algorithms for super-resolution imaging. *IEEE transactions on medical imaging*, 43(8):2970–2987, 2024.

Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11365–11374, 2024.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18278–18289, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Zhiyu Lin, Yifei Gao, and Jitao Sang. Investigating and explaining the frequency bias in image classification. *arXiv preprint arXiv:2205.03154*, 2022.

Heng Liu, Jianyong Liu, Shudong Hou, Tao Tao, and Jungong Han. Perception consistency ultrasound image super-resolution via self-supervised cyclegan. *Neural Computing and Applications*, 35(17):12331–12341, 2023a.

Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2359–2368, 2020.

Tianbao Liu, Zefeiyun Chen, Qingyuan Li, Yusi Wang, Ke Zhou, Weijie Xie, Yuxin Fang, Kaiyi Zheng, Zhanpeng Zhao, Side Liu, et al. Mda-sr: multi-level domain adaptation super-resolution for wireless capsule endoscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 518–527. Springer, 2023b.

Tianyu Liu, Shuai Han, Linru Xie, Wenyu Xing, Chengcheng Liu, Boyi Li, and Dean Ta. Super-resolution reconstruction of ultrasound image using a modified diffusion model. *Physics in Medicine & Biology*, 69(12):125026, 2024.

Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2279–2288, 2025.

Juan Molina, Mircea Petrache, Francisco Sahli Costabal, and Matías Courdurier. Understanding the dynamics of the frequency bias in neural networks. *arXiv preprint arXiv:2405.14957*, 2024.

Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pp. 191–207. Springer, 2020.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 2017.

Yaguan Qian, Renhui Tao, Huabin Du, and Bin Wang. Unveiling the veil: high-frequency components as the key to understanding medical dnns' vulnerability to adversarial examples. *Cybersecurity*, 8(1):21, 2025.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Yuchuan Tian, Hanting Chen, Chao Xu, and Yunhe Wang. Image processing gnn: Breaking rigidity in super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24117, 2024.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8): e0200412, 2018.

Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022.

Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2856–2865, 2024.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018a.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018b.

John Zhi-Qin, John Xu, Luo Tao, Xiao Yanyang, and Ma Zheng. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5): 1746–1767, 2020.

A Related Works

A.1 Single Image Super-Resolution

DL-based approaches have markedly advanced single image super-resolution, surpassing traditional interpolation- and reconstruction-based approaches. Early CNN-based methods such as SRCNN (Dong et al., 2014) optimized networks end-to-end using MSE. Subsequent architectures improved high-frequency recovery through residual (Kim et al., 2016; Lim et al., 2017), dense (Zhang et al., 2018b), multi-scale (Gao & Zhuang, 2019), and attention-based designs (Zhang et al., 2018a). With the rise of Vision Transformers (Dosovitskiy et al., 2020), attention mechanisms were introduced into image restoration. SwinIR (Liang et al., 2021) used shifted window attention to balance local and global features. ART (Zhang et al., 2022) applied sparse attention to reduce complexity, while ATD (Zhang et al., 2024) and PFT (Long et al., 2025) enhanced global modeling through adaptive tokens and progressive attention.

A.2 Medical Image Super-Resolution

SR techniques have been widely applied in medical imaging to enhance visual quality across modalities. Lerendegui et al. (2024) employed localization and tracking to improve microvascular ultrasound imaging for improved diagnosis. Liu et al. (2023a) applied deep learning to achieve high-quality super-resolution under limited ultrasound data. Similarly, Almalioglu et al. (2020) combined conditional adversarial networks with spatial attention to improve endoscopic image resolution. Moreover, Liu et al. (2023b) explored domain-adaptive super-resolution to enhance cross-dataset generalization in endoscopy. Recently, Wavelet-integrated diffusion models were advanced by Liu et al. (2024) to improve the clarity of fetal head ultrasound texture.

A.3 Frequency Domain Methods in Super-Resolution

Current frequency-domain SR methods predominantly utilize wavelet (Huang et al., 2017; Jeevan et al., 2024; Korkmaz et al., 2024; Liu et al., 2024) and Fourier transforms (Fuoli et al., 2021) to improve spectral recovery. Fourier-based approaches, such as Fuoli et al. (2021), apply reconstruction losses in the Fourier domain to suppress high-frequency distortions and enhance detail recovery of SISR. On the other hand, wavelet-based methods decompose images into sub-bands, facilitating detailed reconstruction. For instance, Korkmaz et al. (2024) used wavelet-domain losses to stabilize GAN-based SR and reduce artifacts. In contrast, Liu et al. (2024) incorporated wavelet priors in diffusion models to mitigate over-smoothing and mode collapse in ultrasound images. By explicitly leveraging frequency information, these techniques enhance SISR reconstruction quality.

B Datasets Details

B.1 HC18 Dataset.

The HC18 (van den Heuvel et al., 2018) dataset is part of the HC18 Grand Challenge on automated fetal head circumference (HC) measurement. It contains a total of 1,334 2D ultrasound images acquired from 551 pregnant women undergoing routine scans at Radboud University Medical Center, Nijmegen, Netherlands, between May 2014 and May 2015. Scans were performed using GE Voluson E8 or Voluson 730 devices under approval of the local ethics committee.

B.2 Kvasir Dataset.

The Kvasir (Pogorelov et al., 2017) dataset comprises 4,000 GI endoscopic images evenly distributed across 8 classes (e.g., anatomical landmarks such as Z-line, pylorus and cecum; pathological findings including polyps, ulcers, esophagitis) with 500 images per category. Images were collected at Vestre Viken Health Trust (Norway) and carefully validated by expert gastroenterologists from both the hospital and the Cancer Registry of Norway. All images were acquired using Olympus ScopeGuide hardware.

B.3 IN-HOUSE DATASET.

The in-house dataset consists of 1,658 endoscopic ultrasound (EUS) images collected from 2011 to 2020 at a collaborating hospital using Olympus ScopeGuide hardware. We ensure that data is strictly partitioned by patient, such that no data from the same patient appears in both training and test sets. The study has been approved by the institutional ethics committee.

C TRAINING DETAILS

We conduct all training and validation experiments on three medical image datasets: HC18 (van den Heuvel et al., 2018), Kvasir (Pogorelov et al., 2017), and an in-house dataset. Each dataset is randomly partitioned into training and validation subsets using an 8:2 ratio within each class, ensuring a balanced distribution across categories. To enhance robustness and generalization, we apply standard data augmentations, including random horizontal/vertical flips and 90° rotations.

All models are trained with a fixed upscaling factor of $\times 4$, where each input LR image is of fixed size 64×64. We use the Adam (Kingma & Ba, 2014) optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay coefficient $\lambda = 1 \times 10^{-4}$. The initial learning rate is set to 2×10^{-4} and is decayed by a factor of 0.5 at 150k, 200k, 225k, and 240k iterations via a MultiStepLR scheduler. The total number of training iterations is 250k, with a batch size of 16. The pixel-wise L1 loss is used as the training objective. We implement all models using PyTorch 2.0.1. All experiments are performed on NVIDIA RTX 4090 GPU.

D METRIC DESCRIPTIONS

D.1 PEAK SIGNAL-TO-NOISE RATIO.

Peak Signal-to-Noise Ratio (PSNR) is a widely used objective metric that quantifies the pixel-wise similarity between a reconstructed image and the ground truth (GT). It is calculated using the mean squared error (MSE) and expressed in decibels (dB). Higher PSNR values indicate better reconstruction quality, especially in terms of preserving overall luminance structure. However, PSNR may not align well with perceptual fidelity, as it is highly sensitive to small pixel-level differences.

D.2 STRUCTURAL SIMILARITY INDEX.

The Structural Similarity Index (SSIM) assesses perceived image quality by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. Unlike PSNR, SSIM considers structural information by jointly evaluating luminance, contrast, and texture similarity over local windows. SSIM ranges from 0 to 1, with higher values indicating greater structural fidelity between the reconstructed and GT images.

D.3 LEARNED PERCEPTUAL IMAGE PATCH SIMILARITY.

LPIPS evaluates perceptual similarity by computing distances between deep features extracted from pretrained neural networks (In this study, we adopt VGG). It correlates more closely with human visual perception compared to PSNR or SSIM. Lower LPIPS values indicate that the reconstructed image is perceptually closer to the ground truth. LPIPS is particularly useful for assessing mid- and high-frequency detail and texture realism.

D.4 ENERGY.

Energy quantifies the overall strength of intensity variations by summing the squared gray-level values of an image. It is defined as:

$$\text{Energy} = \sum_{i,j} I(i,j)^2, \tag{9}$$

where $I(i,j)$ denotes the pixel intensity at location (i,j). Higher energy values suggest richer detail content and are often associated with the presence of high-frequency textures.

D.5 SECOND MOMENT.

Second Moment measures the dispersion of pixel intensities around the mean, effectively characterizing the variance of local textures. It emphasizes fluctuations at mid-level scales, making it suitable for analyzing texture complexity beyond uniform regions:

$$\text{Second Moment} = \frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^2, \tag{10}$$

where μ is the mean intensity and N is the number of pixels.

D.6 ENTROPY.

Entropy quantifies the randomness or unpredictability of pixel intensities in an image. It reflects the complexity and richness of textural patterns across the entire frequency spectrum:

$$\text{Entropy} = -\sum_{k} p_k \log(p_k), \tag{11}$$

where p_k is the probability of intensity level k. Higher entropy indicates more diverse gray-level distributions.

D.7 GLCM CONTRAST.

Gray-Level Co-occurrence Matrix (GLCM) Contrast captures the intensity contrast between a pixel and its neighbor over the entire image. It is sensitive to abrupt local changes and is widely used to describe texture roughness:

$$\text{Contrast} = \sum_{i,j}(i-j)^2 P(i,j), \tag{12}$$

where $P(i,j)$ is the normalized GLCM. Larger contrast values indicate more pronounced textural transitions.

D.8 RESIDUAL ENERGY.

Residual Energy (RE) quantifies the overall deviation between the normalized Fourier spectra of the reconstructed image I_{rec} and ground truth I_{GT}, defined as:

$$\text{RE} = \left\| \mathcal{F}(I_{\text{rec}}) - \mathcal{F}(I_{\text{GT}}) \right\|_2, \tag{13}$$

where $\mathcal{F}(\cdot)$ denotes the normalized amplitude spectrum. A lower RE indicates better alignment in spectral content.

D.9 SPECTRAL CORRELATION COEFFICIENT.

Spectral Correlation Coefficient (SCC) measures the linear correlation between the frequency spectra of I_{rec} and I_{GT}, reflecting their structural consistency in the frequency domain:

$$\text{SCC} = \frac{\text{Cov}(\mathcal{F}(I_{\text{rec}}), \mathcal{F}(I_{\text{GT}}))}{\sigma_{\text{rec}}\sigma_{\text{GT}}}, \tag{14}$$

where $\text{Cov}(\cdot,\cdot)$ denotes covariance, and σ_{rec}, σ_{GT} are the standard deviations of the respective spectra. SCC ranges from -1 to 1, with higher values indicating stronger spectral similarity.

D.10 KL DIVERGENCE.

KL Divergence (KLD) evaluates the asymmetric divergence between the spectra of I_{rec} and I_{GT}, interpreting the latter as a target distribution:

$$\text{KLD} = \sum_{k} \mathcal{F}(I_{\text{GT}})_k \log\left(\frac{\mathcal{F}(I_{\text{GT}})_k}{\mathcal{F}(I_{\text{rec}})_k + \epsilon}\right), \tag{15}$$

where k indexes the frequency components and ϵ is a small constant for numerical stability. Lower KLD values indicate better spectral consistency.

Table 5: Extended ablation study on four baseline models (RCAN, SwinIR, ATD, PFT) under the **blind** super-resolution setting. The table shows the effect of incorporating HFSA, AMFD, and FusionGate (FG) modules on three medical datasets. The best results for each group are highlighted in **bold**. We report results using PSNR (↑), SSIM (↑), and LPIPS (↓) as evaluation metrics.

Baseline	Modules			In-house			Kvasir			HC18		
	HFSA	AMFD	FG	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
RCAN Zhang et al. (2018a)	–	–	–	27.66	0.7052	0.1559	30.48	0.8542	0.1287	27.03	0.7665	0.2018
	✓	–	–	27.67	0.7063	0.1559	30.48	0.8545	0.1263	27.17	0.7666	0.2005
	–	✓	–	27.79	**0.7118**	0.1536	31.05	0.8546	0.1283	27.11	0.7668	0.1819
	✓	✓	✓	**27.98**	**0.7118**	**0.1525**	**31.07**	**0.8548**	**0.1222**	**27.33**	**0.7671**	**0.1801**
SwinIR Liang et al. (2021)	–	–	–	29.33	0.7490	0.1386	34.46	0.8753	0.1112	26.31	0.7209	0.1629
	✓	–	–	29.35	0.7493	0.1365	34.53	**0.8755**	0.1112	26.40	0.7256	0.1601
	–	✓	–	29.38	0.7497	0.1361	34.51	0.8754	0.1108	26.33	0.7244	0.1589
	✓	✓	✓	**29.44**	**0.7516**	**0.1353**	**34.55**	**0.8755**	**0.1102**	**26.44**	**0.7330**	**0.1578**
ATD Zhang et al. (2024)	–	–	–	29.40	0.7484	0.1412	34.61	0.8764	0.1108	28.52	0.7959	0.1744
	✓	–	–	29.42	0.7492	0.1402	34.67	0.8770	0.1108	28.70	0.7994	0.1701
	–	✓	–	29.44	0.7502	0.1392	34.77	**0.8774**	0.1108	28.75	0.7997	0.1703
	✓	✓	✓	**29.54**	**0.7541**	**0.1359**	**34.81**	**0.8774**	**0.1105**	**28.76**	**0.8012**	**0.1669**
PFT Long et al. (2025)	–	–	–	29.46	0.7517	0.1380	34.53	0.8762	0.1102	27.94	0.7697	0.1116
	✓	–	–	**29.49**	0.7523	0.1376	**34.55**	0.8763	0.1102	27.95	0.7706	0.1115
	–	✓	–	29.47	0.7518	0.1380	34.54	0.8763	**0.1101**	27.98	0.7708	0.1116
	✓	✓	✓	**29.49**	**0.7527**	**0.1371**	**34.55**	**0.8764**	**0.1101**	**28.03**	**0.7716**	**0.1105**

D.11 Mid- and High-Frequency Energy Ratio.

Mid- and High-Frequency Energy Ratio (MHER) measures the relative energy distributed outside the low-frequency region, offering insight into the model's ability to reconstruct mid- and high-frequency details:

$$\text{MHER} = \frac{\sum_{k \in \Omega_{\text{mid\&high}}} \left(\mathcal{F}(I_{\text{rec}})_k\right)^2}{\sum_k \left(\mathcal{F}(I_{\text{rec}})_k\right)^2}, \tag{16}$$

where $\Omega_{\text{mid\&high}}$ is a frequency band mask that excludes the central region with radial distance less than $0.25 \cdot \min(H, W)$. A higher MHER indicates a stronger ability to reconstruct mid- and high-frequency components.

E Extended Ablation Study on Other Backbones

To further validate the generalizability of our FARE module, we conduct additional ablation experiments under the blind super-resolution setting using four representative baseline models: RCAN, SwinIR, ATD, and PFT. All experiments follow the same protocol as the main paper and are evaluated on the in-house, Kvasir, and HC18 datasets.

As shown in Table 5, consistent with the results observed in the main paper, both HFSA and AMFD individually lead to improvements over their respective baseline models, demonstrating their independent effectiveness. Moreover, integrating both modules into FARE consistently yields the best results across all three metrics, verifying the complementarity of mid- and high-frequency enhancement pathways and the benefits of the proposed cross-frequency fusion mechanism.

F Additional Visual Results

To further validate the effectiveness of the proposed FARE module in recovering semantically meaningful structures and perceptual details, we provide more visual comparisons in the blind super-resolution setting across the HC18, Kvasir, and in-house datasets. As shown in Figure 6, our method consistently restores sharper anatomical boundaries and richer textures than baseline models. Notably, FARE enhances both global fidelity and localized high-frequency details, yielding reconstructions that are more faithful to the ground-truth reference images. These qualitative improvements align well with the quantitative gains reported in Table 5, highlighting the robustness and generalizability of our frequency-aware enhancement strategy.

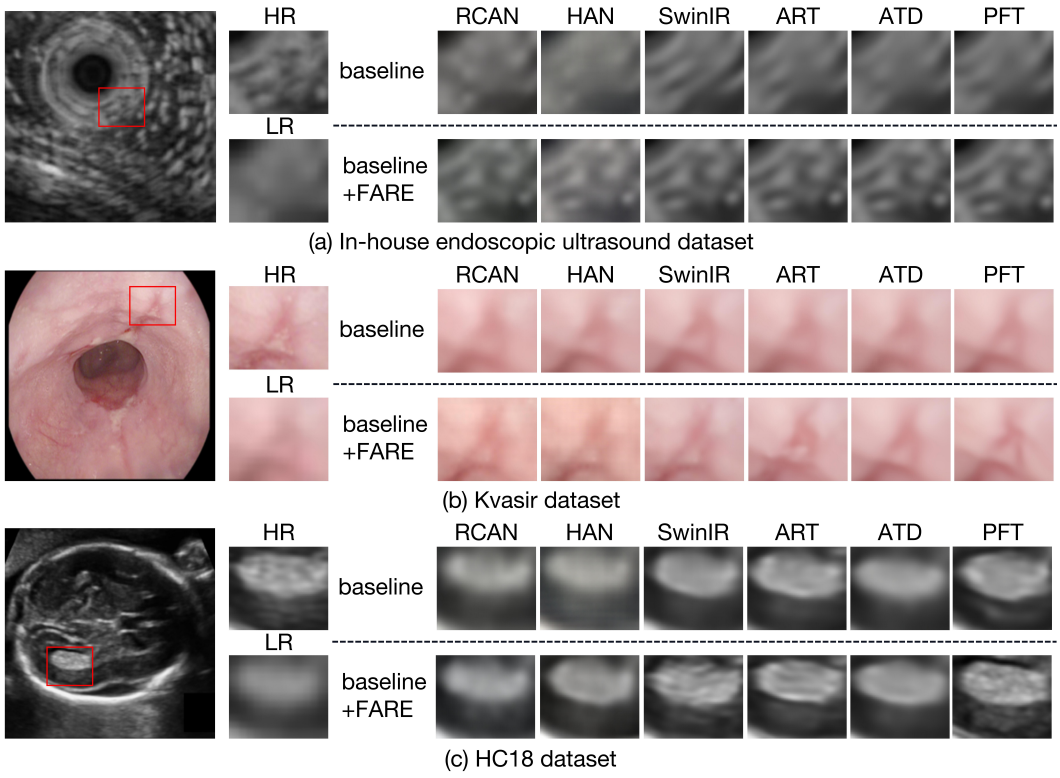


Figure 6: Visual comparison of SR reconstruction results.

G Additional Rationale and Empirical Analysis

G.1 Rationale for Mid- and High-Frequency Decomposition

The proposed frequency decomposition is motivated by the spectral-bias and F-Principle literature, which shows that deep neural networks preferentially learn low-frequency functions and are less efficient at modeling local high-frequency variations (Rahaman et al., 2019; Zhi-Qin et al., 2020). Accordingly, FARE does not replace or weaken the modeling of low-frequency content. Low-frequency information is preserved by the original SR backbone and residual reconstruction pathway, while FARE explicitly complements the backbone by modeling mid- and high-frequency cues that are critical for recovering fine anatomical structures.

The two branches are designed according to the distinct structural roles of these frequency components. HFSA employs DTCWT to capture directional high-frequency discontinuities, such as sharp anatomical boundaries and fine edges. AMFD instead learns a Gaussian bandpass mask to isolate and reconstruct mid-frequency contextual patterns, including tissue textures and gradual anatomical variations. In this sense, "semantic details" refer to frequency-dependent structural roles: mid-frequency responses mainly describe contextual tissue textures and gradual anatomical patterns, whereas high-frequency responses represent sharp boundaries, edges, and local discontinuities. This interpretation is consistent with frequency-aware medical imaging studies (Azad et al., 2021; Bo et al., 2025) and recent wavelet-spectrum SR methods (Du et al., 2025).

G.2 Redundancy and Adaptive Frequency Separation

To quantify the overlap between the extracted mid- and high-frequency representations, we compute normalized mutual information (NMI) between the DTCWT-derived HF responses and Gaussian-bandpass MF responses on the in-house dataset. The resulting NMI is only $\mathbf{0.0747 \pm 0.0016}$, with a redundancy ratio of $\mathbf{0.0756 \pm 0.0028}$. These results indicate limited overlap and little redundant in-

Table 6: Computational overhead of integrating FARE into ATD. Latency is measured for one input image under the same hardware and evaluation setting.

Setting	Params	FLOPs	Latency
ATD (Zhang et al., 2024)	20.3M	1.52T	65ms
ATD+FARE	32.2M	2.06T	96ms

formation between the extracted MF and HF responses, thereby reducing mutual interference during frequency-specific modeling.

Importantly, the proposed decomposition is *soft and adaptive*, rather than a strict hand-crafted partition. AMFD predicts a continuous Gaussian mask conditioned on the input features, allowing the selected mid-frequency range to adapt across images and patches. Residual overlap between MF and HF cues is therefore permitted rather than forcibly discarded. FusionGate is designed to negotiate this residual overlap and complementary information through bidirectional gating, preventing either frequency pathway from dominating the reconstructed representation.

H Additional Computational and Application Analysis

H.1 Computational Overhead

FARE is designed as a plug-in feature enhancer that can be inserted into different SR backbones without modifying their internal blocks, reconstruction heads, or training objectives. "Plug-and-play" and "architecture-agnostic" therefore refer to compatibility with different backbones rather than zero computational cost. Table 6 reports the additional cost when FARE is integrated into ATD.

The integration introduces an additional latency of 31 ms per image. Since the method targets medical image quality enhancement for clinicians and downstream computer-aided diagnosis rather than real-time clinical interaction, this overhead remains acceptable for diagnostic-assistance and preprocessing workflows. The additional cost provides consistent improvements in reconstruction fidelity and structural detail preservation.

H.2 Applicability and Medical Relevance

The experiments cover multiple imaging modalities and anatomical regions, including endoscopic ultrasound, fetal ultrasound, gastrointestinal endoscopy, and BraTS/IXI MRI, as well as blind and non-blind degradation settings. The consistent improvements across these datasets support the broad applicability of FARE across modalities, organs, and degradation types.

For endoscopic and fetal ultrasound, clearer anatomical boundaries can assist lesion-layer localization, invasion-range assessment, tissue-boundary interpretation, and the extraction of more reliable features for computer-aided diagnosis. Thus, the role of SR in these scenarios is not limited to improving visual appearance; it also provides higher-quality inputs for subsequent lesion recognition, classification, and diagnostic-assistance systems.

The MRI experiments serve as an additional transferability evaluation under the established multi-contrast MRI super-resolution paradigm rather than as a replacement for the SISR setting. FARE is inserted into the intra-network feature blocks while the original cross-contrast fusion mechanism remains unchanged, allowing the experiment to assess whether frequency-aware enhancement complements cross-contrast guidance.

I Additional Transform Ablation and Backbone Evaluation

The main paper reports the structure-aware prior ablation, where setting the gradient-guidance tensor to all ones (grad=1) degrades HFSA. This result validates that the gradient prior is necessary for boundary-aware high-frequency enhancement. We further evaluate the choice of wavelet transform and the compatibility of FARE with additional recent SR backbones.

Table 7: Additional results under the blind $\times 4$ super-resolution setting. The first group compares DTCWT with Haar wavelets in ATD+FARE. The remaining groups evaluate FARE on the additional IPG, HAT, and ESC backbones. The best result within each group is highlighted in **bold**.

Backbone	Configuration	In-house			Kvasir		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ATD (Zhang et al., 2024)	Haar wavelet	29.42	0.7523	0.1387	34.65	0.8767	0.1108
	DTCWT	**29.54**	**0.7541**	**0.1359**	**34.81**	**0.8774**	**0.1105**
IPG (Tian et al., 2024)	Base	29.33	0.7477	0.1459	34.52	0.8755	0.1110
	+FARE	**29.40**	**0.7503**	**0.1375**	**34.66**	**0.8763**	**0.1106**
HAT (Chen et al., 2023)	Base	29.28	0.7479	0.1436	34.20	0.8743	0.1117
	+FARE	**29.37**	**0.7511**	**0.1378**	**34.45**	**0.8769**	**0.1112**
ESC (Lee et al., 2025)	Base	28.26	0.7370	0.1564	33.98	0.8673	0.1181
	+FARE	**28.91**	**0.7457**	**0.1406**	**34.33**	**0.8728**	**0.1142**

Replacing DTCWT with Haar wavelets consistently reduces performance on both datasets. DTCWT provides directional subbands and approximate shift invariance, which are better suited to representing anatomical boundaries and local structural discontinuities. The result therefore supports the use of DTCWT rather than a generic wavelet decomposition in HFSA.

FARE also consistently improves IPG (Tian et al., 2024), HAT (Chen et al., 2023), and ESC (Lee et al., 2025) in PSNR, SSIM, and LPIPS. These results complement the RCAN, SwinIR, ATD, and PFT experiments reported above, showing that the effectiveness of FARE is not tied to a particular convolutional or Transformer backbone. Instead, the gains remain consistent across architectures with different feature-extraction and attention mechanisms.

Among the recent SR methods examined in this study, ESC was the method for which the released training and testing protocol could be directly reused within our feed-forward medical SR setting. We therefore conduct an additional controlled evaluation on ESC. As shown in Table 7, ESC+FARE improves all three metrics on both datasets, further validating the compatibility of FARE with recent efficient SR architectures.

J Statistical Significance on PFT

To determine whether the relatively small average improvement over PFT could be attributed to normal performance variation, we conduct paired per-image Wilcoxon signed-rank tests on PSNR, SSIM, and LPIPS using the same test images under the in-house blind-SR setting. Signed improvements are defined as PFT+FARE minus PFT for PSNR and SSIM, and PFT minus PFT+FARE for LPIPS, such that positive values consistently indicate improvement. All three metrics satisfy $p < 0.05$, confirming that the gains are statistically significant rather than random fluctuations.

The visual comparisons in Figure 3 of the main paper and Figure 6 further support the statistical results. Although the average numerical gain over the strong PFT backbone is relatively modest, FARE recovers clearer anatomical textures and sharper local structures, indicating improvements in perceptual quality and structural fidelity.

K Additional Implementation and Ethics Clarifications

The in-house study received approval from the institutional ethics committee under IRB No. 2019-SR-448 and was conducted in accordance with the principles of the Declaration of Helsinki. The dataset was partitioned strictly by patient identity so that images from the same patient did not appear in both the training and test sets.