Versatile and Efficient Medical Image Super-Resolution via Frequency-Gated Mamba

Wenfeng Huang[1,2], Xiangyun Liao[2], Wei Cao[3], Wenjing Jia[1], and Weixin Si[4*]
[1]*Faculty of Engineering and Information Technology, University of Technology Sydney, Australia*
[2]*Shenzhen Institutes of Advanced Technology, China*
[3]*College of Computer Science and Technology, Qingdao University, China*
[4]*Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, China*

***Abstract*—Medical image super-resolution (SR) is essential for enhancing diagnostic accuracy while reducing acquisition cost and scanning time. However, modeling both long-range anatomical structures and fine-grained frequency details with low computational overhead remains challenging. We propose FGMamba, a novel frequency-aware gated state-space model that unifies global dependency modeling and fine-detail enhancement into a lightweight architecture. Our method introduces two key innovations: a Gated Attention-enhanced State-Space Module (GASM) that integrates efficient state-space modeling with dual-branch spatial and channel attention, and a Pyramid Frequency Fusion Module (PFFM) that captures high-frequency details across multiple resolutions via FFT-guided fusion. Extensive evaluations across five medical imaging modalities (Ultrasound, OCT, MRI, CT, and Endoscopic) demonstrate that FGMamba achieves superior PSNR/SSIM while maintaining a compact parameter footprint (<0.75M), outperforming CNN-based and Transformer-based SOTAs. Our results validate the effectiveness of frequency-aware state-space modeling for scalable and accurate medical image enhancement. Source code and dataset will be made publicly available.**

***Index Terms*—Medical Image, Super-Resolution, Mamba, State-Space Model, Lightweight Network**

I. Introduction

High-resolution medical imaging plays a vital role in accurate clinical diagnosis and treatment planning. However, acquiring high-quality images—especially high-resolution magnetic resonance imaging (MRI)—often requires lengthy scan times and expensive hardware, imposing practical constraints in routine care. Super-resolution (SR) techniques aim to address this limitation by reconstructing high-resolution images from low-resolution acquisitions, thereby enabling detailed anatomical and pathological visualization at reduced cost and time.

The emergence of convolutional neural networks (CNNs) marked a paradigm shift in SR research by enabling end-to-end learning of nonlinear mappings from LR to HR images. SRCNN [1] pioneered this approach using a shallow network, which was soon surpassed by deeper and more expressive architectures such as EDSR [2], CARN [3], and LapSRN [4], which incorporated residual learning and multi-scale feature fusion. These architectures laid the foundation for a wide array of variants, including frequency-aware designs such as CFSRCNN [5], which models coarse-to-fine representations, and LESRCNN [6], which integrates sub-pixel convolutions with dense blocks for improved efficiency.

In the domain of medical imaging, CNN-based SR models have been widely adopted and extended to accommodate various imaging modalities and clinical requirements. Dual U-Net residual architectures [7] and volumetric 3D CNNs [8] have been leveraged for high-fidelity MRI and cardiac image restoration, preserving anatomical coherence across spatial slices. Generative adversarial networks (GANs) have also proven effective in capturing perceptual realism, as demonstrated in the progressive GAN strategies for MRI and retinal imaging [9]. Furthermore, inverse-consistent GANs [10] were developed to ensure structural symmetry in OCT super-resolution, while Goyal et al. [11] utilized multi-scale cascaded CNNs for ultrasound image enhancement.

To better address frequency-aware signal modeling and modality-specific constraints, recent research introduced attention-based mechanisms and frequency-domain priors. Mix-attention architectures effectively integrate spatial and channel attention in pathological image SR, while some research [12], [13] explicitly exploits the Fourier Transform to enhance frequency representation in MRI sequences. These advances reflect the growing recognition of hybrid spatial-frequency modeling as critical for clinical-grade SR performance.

Despite the empirical success of CNN-based techniques, their local convolutional kernels fundamentally limit their ability to model long-distance relationships and global structural patterns. This constraint becomes particularly problematic in high-resolution 3D medical data, where anatomical consistency must be preserved across large spatial extents.

To overcome the locality limitation of convolutional kernels, Vision Transformers (ViTs) and hybrid CNN–Transformer architectures have been introduced to better model long-distance relationships via global attention mechanisms. Among them, SwinIR [14] employs shifted window-based self-attention combined with residual connections and hierarchical representations, striking a balance between global context modeling and computational efficiency. SwinIR and its medical adaptations have shown strong performance across modalities such as

* Corresponding author.
Wenfeng Huang and Xiangyun Liao are co-first authors of this paper.

MRI and endoscopic imaging, often achieving higher PSNR and SSIM scores than CNN-based counterparts, while also improving perceptual and diagnostic quality [15]. In particular, SwinIR's ability to capture multi-scale structural priors and semantic coherence has proven beneficial in enhancing subtle anatomical features.

In parallel, other Transformer-inspired models have further enriched the SR landscape. ESRT [16], for instance, combines convolutional and Transformer branches through early fusion and skip connections, achieving strong results with relatively low parameter cost. Some researchs [12], [13], [17] combine transformer and frequency information for medical image super-resolution.

In the medical domain, these models have been increasingly applied across a range of modalities—ultrasound, MRI, CT, endoscopic, and OCT—each with distinct noise characteristics and anatomical priors. To better adapt to these challenges, hybrid frameworks such as LGSR [15] have emerged. LGSR integrates a local-to-global feature learning pipeline that fuses windowed attention with sparse token selection, enabling efficient contextual interaction while maintaining lightweight design. It demonstrates state-of-the-art performance across ultrasound, OCT, and MRI datasets, offering robust anatomical consistency and high-frequency restoration.

Recently, structured state-space models [18]–[21] (SSMs), and in particular the emergent class of Mamba-style architectures, have shown remarkable efficiency in modeling long-range dependencies with linear complexity. Mamba models, offering selective scanning mechanisms and hardware-aware optimizations, yielding strong sequence modeling power with greatly reduced parameter and memory overheads . The adaptation of Mamba to low-level vision tasks via MambaIR [22], which combine convolutional layers and channel attention with state-space modules to address local spatial recurrences and channel redundancy in restoration tasks. Such researchs [23], [24] has been proven the success of Mamba in image restoration tasks.

Although Transformer-based super-resolution models such as LGSR [15] have advanced the state of the art in medical image enhancement, their reliance on self-attention mechanisms inevitably introduces quadratic complexity, which limits scalability for high-resolution or volumetric medical data. While CNN-based approaches remain computationally efficient, they struggle to model global dependencies critical for anatomical consistency, especially in modalities such as MRI or OCT where context-aware reconstruction is essential.

To bridge this gap, we propose FGMamba, a novel frequency-aware state-space framework that integrates gated attention and multiscale frequency residual learning for efficient and accurate medical image super-resolution. Inspired by recent advances in state-space sequence modeling, our method captures long-range dependencies with linear complexity while maintaining lightweight design. Unlike prior approaches that rely on windowed or token-based attention, our model introduces a frequency-guided residual feedback mechanism that explicitly enhances high-frequency details—key to restoring structural sharpness in degraded medical scans. Additionally, the gated attention unit selectively enhances discriminative spatial-channel information, further improving texture fidelity and edge continuity. Our main contributions are summarized as follows:

- We design a Pyramid Frequency Fusion Module (PFFM) that explicitly enhances high-frequency details by decomposing feature maps across multi-scale FFT domains. This module guides the reconstruction process to recover sharp anatomical boundaries and texture details essential in clinical diagnosis.
- We introduce a Gated Attention-enhanced State-Space Module (GASM) that augments the vanilla VSSM2D block with spatial and channel attention units. This hybrid design allows selective emphasis on discriminative features while maintaining the memory and runtime efficiency of structured state-space modeling.
- Extensive experiments across five benchmark medical imaging datasets (ultrasound, OCT, MRI, CT, endoscopic) show that our model outperforms existing CNN-, Transformer-, and Mamba-based SR methods in both PSNR/SSIM and qualitative fidelity—despite using fewer than 0.75M parameters.

II. Related Works

A. Image Super-Resolution

Single-image super-resolution (SISR) has evolved through several phases. Early approaches were grounded in mathematical and statistical techniques such as Random Forest regression, anchored neighborhood regression [31], and dictionary learning, yet they lacked the adaptability to recover complex visual structures. The advent of convolutional neural networks (CNNs) revolutionized SISR: SRCNN [1] introduced end-to-end mapping from LR to HR images, and deeper models like EDSR [2], CARN [3], and LapSRN [4] incorporated residual and pyramid architectures to enhance restoration quality. CFSRCNN [5] and LESRCNN [6] further emphasized frequency-aware and lightweight design through dense residual connections and sub-pixel components.

In medical imaging, the adaptation of CNN-based SR has shown strong progress. Dual U-Net residual [7] structures and 3D residual CNNs have been used for cardiac MRI restoration, improving local detail preservation and anatomical coherence. GAN-based methods also emerged: Mahapatra et al. employed progressive GANs for retinal and MRI SR [9]. OCT super-resolution [32] has been tried via inverse-consistent GANs [10], self-supervised learning [32], and transformer [33], and ultrasound SR [34] has benefited from semi-supervised GAN [35]. Despite these advances, CNNs remain fundamentally constrained by their local receptive fields, limiting global consistency—especially in large volumetric datasets.

To address this limitation, hybrid CNN-Transformer architectures were proposed. Such as LGSR [15]—a CNN-ViT model designed for medical image SR that combines deformable CNN layers and global Transformers to learn both

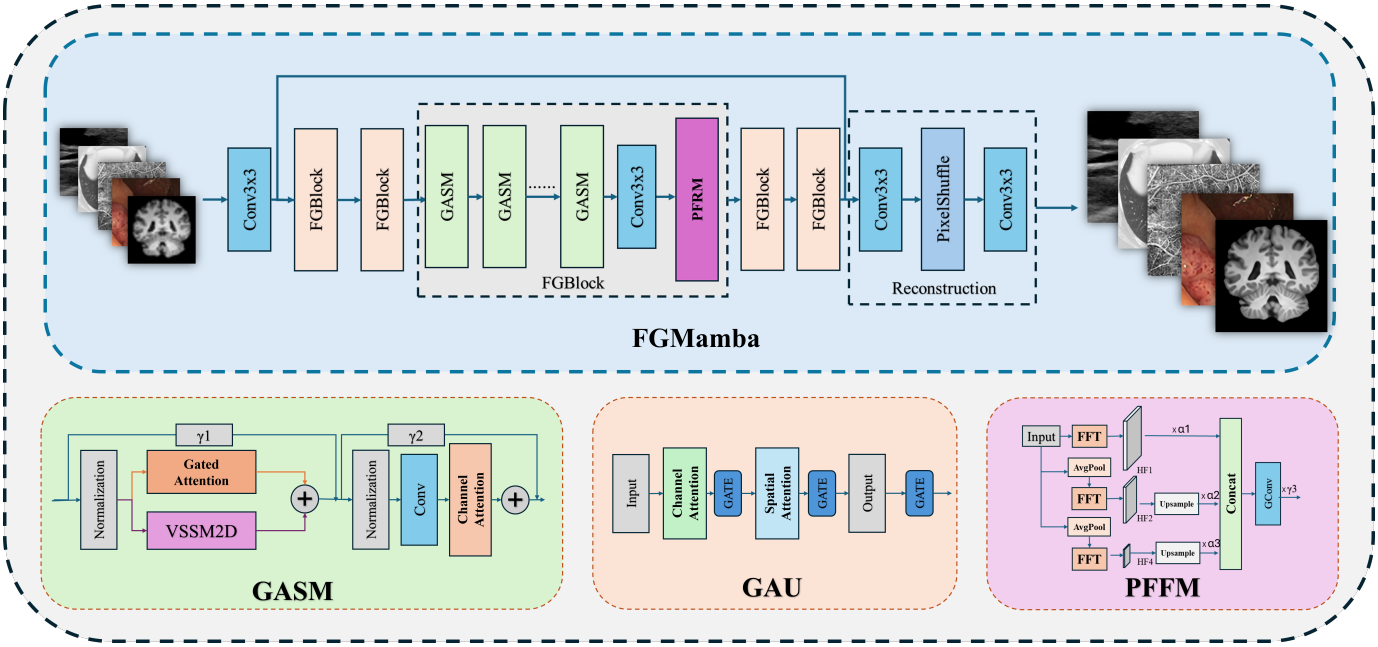

Fig. 1. Overall architecture of the proposed **FGMamba**. It consists of an initial convolution, several FGBlocks, and a reconstruction module with pixel-shuffle upsampling. Each FGBlock contains multiple GASM (Gated Attention State Space Modules), a frequency-enhancing PFFM (Pyramid Frequency Fusion Module), and additional Mamba residual connections. The submodules are illustrated below: (1) GASM incorporates VSSM2D with gated spatial/channel attention, (2) GAU enhances feature selection via dual attention gating, and (3) PFFM extracts and fuses high-frequency components across multiple scales via FFT-based filtering and residual learning.

TABLE I
QUANTITATIVE COMPARISON ON **ULTRASOUND** DATASET AT $\times 2$ SCALE. BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Architecture	Dataset	Scale	Parameters	PSNR (dB)	SSIM
SRCNN [1]	CNN	Ultrasound [15]	2x	69K	37.39	0.9400
CARN [3]				1.59M	37.68	0.9440
EDSR-baseline [2]				1.37M	37.72	0.9447
CFSRCNN [5]				1.49M	37.62	0.9433
LESRCNN [6]				0.81M	37.47	0.9428
ESRT [16]	CNN+ViT			0.68M	37.61	0.9436
LBNET [25]				0.73M	37.51	0.9418
LGSR [15]				0.90M	37.73	0.9448
FGMamba	**CNN+Mamba**			**0.72M**	**38.13**	**0.9511**

local detail and long-range semantic context. Such method achieves superior PSNR/SSIM across multiple modalities (Ultrasound, OCT, CT, MRI). However, methods with ViTs still cost a lot because of their self-attention mechanisms.

B. Mamba Architectures

Mamba [18], [19] is a recently proposed state-space sequence modeling architecture that offers a compelling alternative to Transformers by addressing their quadratic complexity bottleneck. As a selective state-space model (SSM) [20], [21], Mamba captures long-range dependencies via continuous-time dynamics while maintaining linear inference complexity, making it highly scalable to high-resolution visual inputs. Unlike traditional RNNs or attention-based mechanisms, Mamba decouples memory access from state updates, allowing dynamic selection of relevant information at each step. This architecture has demonstrated remarkable efficiency in sequence modeling and is now rapidly gaining traction in vision tasks.

The initial success of Mamba in language modeling has spurred several adaptations to low-level vision problems. For example, MambaIR [22] introduces a residual structure that combines spatial encoding and global receptive fields via 2D Mamba modules. These developments [36], [37] demonstrate that Mamba-based designs are not only parameter-efficient but also well-suited for tasks requiring both local detail enhancement and global structural understanding. The intrinsic ability of Mamba to model long-term dependencies with low memory overhead makes it an attractive backbone for super-resolution architectures, particularly in the medical domain, where high-resolution volumetric data pose significant

TABLE II
QUANTITATIVE COMPARISON ON **ULTRASOUND** DATASET AT $\times 3$ SCALE. BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Architecture	Dataset	Scale	Parameters	PSNR (dB)	SSIM
SRCNN [1]	CNN	Ultrasound [15]	3x	69K	33.53	0.8576
CARN [3]				1.59M	33.66	0.8612
EDSR-baseline [2]				1.55M	33.71	0.8619
CFSRCNN [5]				1.54M	33.54	0.8591
LESRCNN [6]				0.81M	33.56	0.8598
ESRT [16]	CNN+ViT			0.77M	33.66	0.8617
LBNET [25]				0.74M	33.63	0.8606
LGSR [15]				0.90M	33.74	0.8622
FGMamba	**CNN+Mamba**			**0.73M**	**33.91**	**0.8659**

TABLE III
QUANTITATIVE COMPARISON ON **MULTI-MODAL DATASETS** AT $\times 4$ SCALE. BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Architecture	Dataset	Scale	Parameters	PSNR (dB)	SSIM
SRCNN [1]	CNN	OCTA-500 [26] (OCT)	4x	69K	20.75	0.4474
		CVC-ClinicDB [27](Endoscopic)			35.65	0.9162
		SARS-COV-2 [28] (CT)			33.96	0.8645
		NFBS [29] (MRI)			27.76	0.8598
VDSR [30]		OCTA-500 [26] (OCT)		0.6M	20.86	0.4603
		CVC-ClinicDB [27] (Endoscopic)			36.68	0.9269
		SARS-COV-2 [28] (CT)			35.01	0.8791
		NFBS [29] (MRI)			29.06	0.8943
Lapsrn [4]		OCTA-500 [26] (OCT)		0.81M	20.89	0.4630
		CVC-ClinicDB [27] (Endoscopic)			36.33	0.9226
		SARS-COV-2 [28] (CT)			35.22	0.8815
		NFBS [29] (MRI)			29.58	0.9049
ESRT [16]	CNN+ViT	OCTA-500 [26] (OCT)		0.75M	20.89	0.4627
		CVC-ClinicDB [27] (Endoscopic)			36.12	0.9251
		SARS-COV-2 [28] (CT)			35.25	0.8815
		NFBS [29] (MRI)			29.38	0.9003
LBNET [25]		OCTA-500 [26] (OCT)		0.74M	20.90	0.4633
		CVC-ClinicDB [27] (Endoscopic)			36.48	0.9238
		SARS-COV-2 [28] (CT)			35.37	0.8823
		NFBS [29](MRI)			29.50	0.9031
LGSR [15]		OCTA-500 [26] (OCT)		0.9M	20.91	0.4635
		CVC-ClinicDB [27] (Endoscopic)			36.88	0.9287
		SARS-COV-2 [28] (CT)			35.44	0.8840
		NFBS [29](MRI)			29.75	0.9084
FGMamba	**CNN+Mamba**	OCTA-500 [26] (OCT)		**0.74M**	**20.98**	**0.4697**
		CVC-ClinicDB [27] (Endoscopic)			**37.32**	**0.9290**
		SARS-COV-2 [28] (CT)			**36.14**	**0.8985**
		NFBS [29](MRI)			**29.91**	**0.9129**

computational challenges. By integrating frequency-aware representation and gated attention mechanisms into a lightweight Mamba backbone, our FGMamba architecture enables comprehensive spatial–spectral feature learning while maintaining low parameter overhead. Unlike conventional Transformer or convolutional designs, FGMamba leverages state space modeling for efficient long-range dependency modeling, while simultaneously enhancing fine-grained detail restoration through residual frequency modulation and selective attention. This makes FGMamba a promising solution for high-resolution medical image enhancement across diverse modalities such as OCT, MRI, and CT.

III. METHOD

Our proposed framework, FGMamba, is a compact yet effective medical image super-resolution model inspired the recent researchs [15], [18], [22]. While we inherit the efficient long-range modeling capacity of Mamba, we introduce two key innovations to enhance detail restoration: a Pyramid Frequency Fusion Module (PFFM) and a Gated Attention(GA). The full architecture is illustrated in Fig. 1. We begin by applying a 3×3 convolutional layer to extract shallow features

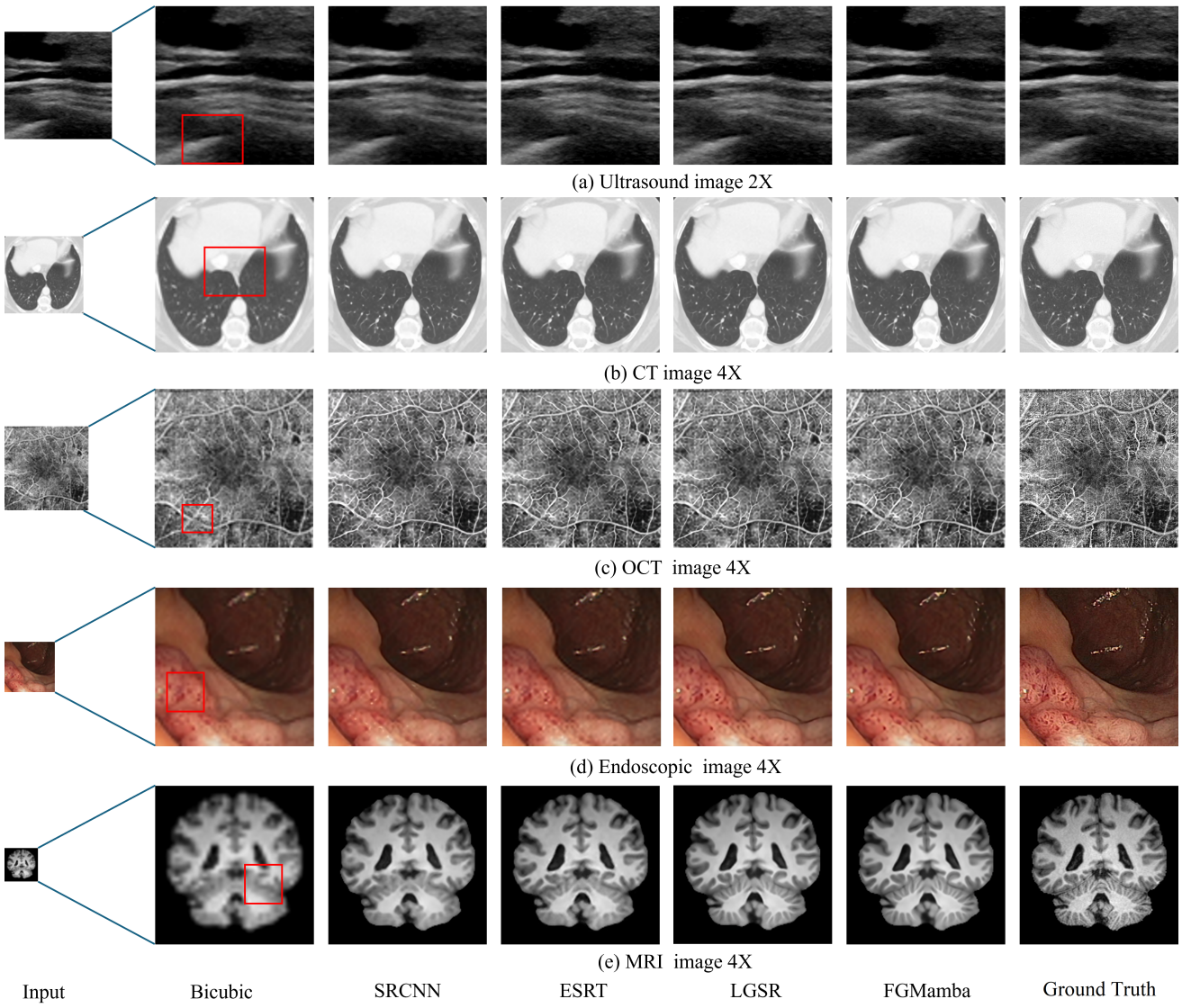

Fig. 2. Visual comparison across five medical modalities: (a) Ultrasound (×2), (b) CT (×4), (c) OCT (×4), (d) Endoscopic (×4), and (e) MRI (×4). Red boxes highlight diagnostically critical structures (vessels, lesions, tissue textures), where FGMamba achieves sharper, more detailed reconstructions to aid radiologist diagnosis and clinical decision-making.

from the low-resolution medical image. Let the input image be denoted as X, then the initial feature map is obtained as:

$$F_x = \text{Conv}_{3\times3}(X), \tag{1}$$

where F_x represents the extracted shallow feature representation.

And then, we proposed FGBlock, which is the core block of our method.

A. FGBlock

The FGBlock serves as the core computation module of FGMamba, composed of two essential components: the Gated Attention-enhanced State-Space Module (GASM) and the Pyramid Frequency Fusion Module (PFFM). This dual-branch structure allows the network to effectively combine long-range dependency modeling with fine-grained high-frequency detail enhancement.

1) Gated Attention-enhanced State-Space Module (GASM): Given the input feature map $F_x \in \mathbb{R}^{C\times H\times W}$, we first apply Layer Normalization:

$$F_{\text{norm}} = \text{Norm}(F_x). \tag{2}$$

Then, the normalized feature is processed through two branches:

A state-space branch using the VSSM2D [22] module:

$$F_{\text{vssm}} = \text{VSSM2D}(F_{\text{norm}}), \tag{3}$$

and a gated attention branch:

$$F_{\text{gate}} = \text{GA}(F_{\text{norm}}), \tag{4}$$

where GA$(\cdot)$ denotes our proposed Gated Attention Unit and GATE$(\cdot)$ is a learnable gate controller. The core of the gated attention branch is our proposed Gated Attention Unit (GAU), which integrates both channel-wise and spatial-wise attention [38] in a lightweight yet effective manner. Specifically, given input $F \in \mathbb{R}^{B\times C\times H\times W}$:

- *Channel Attention*: A global context descriptor is obtained via adaptive average pooling, followed by two fully connected layers to compute a channel-wise attention map $A_c \in \mathbb{R}^{B\times C\times 1\times 1}$.

- *Spatial Attention*: The input is aggregated across the channel dimension using both average pooling and max pooling. The concatenated result is passed through a $k \times k$ convolution (default k=7) to yield a spatial attention map $A_s \in \mathbb{R}^{B\times 1\times H\times W}$.

The final attended feature is computed as:

$$F_{\text{gate}} = F \odot (A_c \odot Gate \odot A_s \odot Gate) \odot Gate, \quad (5)$$

where $\odot$ denotes element-wise multiplication. And the $Gates$ are learned by $Sigmoids$. This gating mechanism enables the network to selectively emphasize semantically informative features while suppressing irrelevant background noise.

The outputs from both branches are fused with a learnable residual scaling parameter γ_1:

$$F_{\text{add}} = F_{\text{vssm}} + F_{\text{gate}} + \gamma_1 \cdot F_x. \quad (6)$$

Then, we pass F_{add} through a convolution and channel attention block:

$$F_{\text{norm2}} = \text{Norm}(F_{\text{add}}), \quad (7)$$

$$F_{\text{conv}} = \text{Conv}_{3\times3}(F_{\text{norm2}}), \quad (8)$$

$$F_{\text{gasm}} = \text{Conv}_{3\times3}(\text{Channel Attention}(F_{\text{conv}}) + \gamma_2 \cdot F_x). \quad (9)$$

2) Pyramid Frequency Fusion Module (PFFM): To extract texture-rich information, we introduce a frequency-domain enhancement module. For each scale $s \in \{1, 2, 4\}$, the input F_{gasm} is downsampled:

$$F_s = \begin{cases} F_{\text{gasm}}, & \text{if } s = 1, \\ \text{AvgPool}(F_{\text{gasm}}, s), & \text{otherwise.} \end{cases} \quad (10)$$

The 2D Fourier Transform is applied:

$$\mathcal{F}_s = \text{FFT}(F_s), \quad (11)$$

followed by a high-frequency mask:

$$M_s = \mathbb{1}(|\mathcal{F}_s| > \mu_{\mathcal{F}_s}), \quad (12)$$

and inverse FFT to get high-frequency spatial features:

$$H_s = \text{IFFT}(\mathcal{F}_s \cdot M_s). \quad (13)$$

Each H_s is upsampled back to original resolution and weighted by a learnable coefficient α_s:

$$F_s^{\uparrow} = \alpha_s \cdot \text{Upsample}(H_s). \quad (14)$$

The fused high-frequency feature is:

$$F_{\text{PFFM}} = \gamma \cdot \text{GroupConv}_{1\times1}\left(\text{Concat}(F_1^{\uparrow}, F_2^{\uparrow}, F_4^{\uparrow})\right), \quad (15)$$

where γ is a learnable scale parameter. The feature representation after passing through multiple FGBlocks is denoted as F_{FGB}.

4) Reconstruction Module: After passing through a series of FGBlocks, the final feature map is denoted as $F_{\text{final}=F_(FGB)+F_x}$. To recover the high-resolution image from this deep representation, we adopt a simple yet effective reconstruction pipeline.

Specifically, the reconstruction block consists of a 3×3 convolution to refine the features, followed by a PixelShuffle operation [39] to upscale the spatial resolution. A final 3×3 convolution layer is then applied to generate the output image:

$$I_{\text{SR}} = \text{Conv}_{3\times3}\left(\text{PixelShuffle}\left(\text{Conv}_{3\times3}(F_{\text{final}})\right)\right), \quad (16)$$

where I_{SR} denotes the super-resolution image.

IV. Experiments

A. Datasets

1) Ultrasound Image Dataset: We use the breast ultrasound dataset [15]. It includes 500 high-resolution scans acquired using GE Vivid Iq and E9 systems in "Breast" mode. After sliding-window cropping and augmentation, we generate 12,000 training and 1,250 testing patches of sizes 240×240 and 256×256, respectively. LR images are created via bicubic downsampling at scales $2\times$, and $3\times$.

2) OCT Image Dataset: We use the "OCTA-6M-Projection Map-OCTA(FULL)" subset from OCTA-500 [26], which contains 300 retinal images (400×400 pixels). For SR tasks, we downsample them to 100×100 via bicubic interpolation and split the dataset into 80% for training and 10% each for validation and testing.

3) Endoscope Image Dataset: Colonoscopy frames are selected from CVC-ClinicDB [27], consisting of 612 images extracted from 31 video sequences. To remove irrelevant dark and overlaid regions, images are cropped to 240×240 and downsampled to 60×60 for paired training. We adopt an 8:1:1 split for training, validation, and testing.

4) CT Image Dataset: We utilize the SARS-CoV-2 CT dataset [28], selecting the 1,230 scans from healthy individuals. After cropping via a sliding window, we obtain 5,432 CT patches at 200×200 resolution. LR versions are generated by bicubic downsampling. The dataset is split into 983 training, 123 validation, and 123 testing samples.

5) MRI Image Dataset: MRI data is sourced from the NFBS repository [29], which provides 125 skull-stripped T1 brain scans. We extract the final 30 slices per scan—regions with fuller anatomical content—and crop each to 160×160. LR images (40×40) are created by downsampling. The split includes 3,000 training, 375 validation, and 375 testing images.

B. Impletement details

Following standard practices in recent literature [15], [22], we augment the training data using horizontal mirroring and random rotations of $90°$, $180°$, and $270°$. For patch-based learning, we divide each image into fixed-size patches, with

TABLE IV
ABLATION STUDIES ON OCT (CT) DATASET AT $\times 4$ SCALE.

Method	PSNR (dB)	SSIM
Baseline (without GAU & PFFM)	20.9750	0.4686
w/o GAT (no GAU in GASM)	20.9765	0.4693
w/o Freq (no PFFM frequency module)	20.9740	0.4684
FGMamba (full model)	**20.9781**	**0.4697**

patch dimensions dynamically adjusted depending on the dataset and upscaling factor. To maintain consistency across experiments, we adopt a batch size of 8. Optimization is conducted using the Adam algorithm with momentum parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized at 2×10^{-4}. All experiments are carried out on a single NVIDIA RTX 4090 GPU.

C. Comparison with the State of the Arts

To evaluate the effectiveness of our proposed FGMamba we conduct comprehensive comparisons against a range of state-of-the-art (SOTA) super-resolution models, including traditional CNN-based methods (SRCNN [1], VDSR [30], LapSRN [4]), Transformer-integrated variants (ESRT [16], LBNET [25]), and recent gated attention approaches (LGSR [15]). Quantitative results across multiple medical image modalities—Ultrasound, CT, OCT, endoscopic, and MRI—are presented in Tables II, II, and III.

As shown, FGMamba consistently achieves superior performance with significantly fewer parameters. For instance, under the challenging $\times 4$ scale across multi-modal datasets (Table III), FGMamba attains the highest PSNR/SSIM scores in all domains, such as 37.32 dB / 0.9290 (endoscopic), 36.14 dB / 0.8985 (CT), and 29.91 dB / 0.9129 (MRI), outperforming the ViT-based LBNET and LGSR models while using fewer parameters (0.74M vs. 0.9M). Similarly, for $\times 3$ scale on ultrasound images (Table II), FGMamba leads with 33.91 dB / 0.8659, demonstrating its robustness in lower-resolution recovery tasks.

The qualitative results shown in Fig. 2 further highlight FGMamba's ability to restore high-frequency textures and preserve anatomical structures across different imaging types. Compared to other methods, FGMamba produces visually sharper boundaries, more realistic textures, and fewer artifacts, especially evident in vascular (OCT) and gastrointestinal (endoscopic) scenes. These improvements not only enhance perceptual fidelity but also provide clearer visualization in diagnostically critical regions, facilitating more accurate lesion detection and anatomical analysis.

D. Ablation Study

To validate the effectiveness of our proposed components, we conduct an ablation study on the OCT dataset at $\times 4$ scale. As shown in Table IV, removing the gated attention mechanism in GASM slightly degrades both PSNR and SSIM, confirming its contribution to structure enhancement. Similarly, excluding the frequency-domain PFFM module results in further performance drops, indicating the importance of multi-scale high-frequency restoration. The full FGMamba model achieves the best results, demonstrating the complementary benefits of both modules.

V. DISCUSSION

As shown in Fig. 2, the red box regions highlight diagnostically important structures such as vessels, lesions, and tissue textures. FGMamba produces sharper and more detailed reconstructions in these areas, which can assist radiologists in improved diagnosis and clinical decision-making. Specifically, our frequency-aware gated state-space architecture generates significantly sharper anatomical boundaries and richer textural details compared to existing SOTAs. Gated Attention-enhanced State-Space Module synergizes dual-branch spatial and channel attention with efficient state-space modeling, while the Pyramid Frequency Fusion Module exploits FFT-guided fusion to captures high-frequency details across multiple resolutions. Clinically, these enhancements directly translate to improved diagnostic confidence in identifying early-stage pathologies and refining treatment planning. Furthermore, the enhanced perceptual quality across five modalities (Ultrasound, OCT, MRI, CT, Endoscopy) facilitates downstream tasks like segmentation or detection, potentially boosting the accuracy of automated analysis. Given its lightweight design ($<$0.75M parameters) and modality-agnostic framework, FGMamba shows strong potential for deployment in clinical systems. In future work, we plan to evaluate its impact on representative downstream tasks to further validate its practical utility.

VI. CONCLUSION

In this paper, we introduced FGMamba, a lightweight and frequency-aware super-resolution framework tailored for medical imaging. By integrating a gated attention mechanism with structured state-space modeling (GASM) and enhancing high-frequency detail via a pyramid frequency fusion module (PFFM), our method effectively captures both global contextual patterns and fine structural cues. FGMamba demonstrates superior PSNR and SSIM performance across five distinct medical modalities while maintaining under 0.75M parameters. Extensive evaluations validate its ability to recover sharp anatomical boundaries and texture details, offering a promising and scalable solution for clinical image enhancement tasks. Future work may explore its extension to volumetric SR and real-time deployment in diagnostic systems.

VII. ACKNOWLEDGEMENTS

This work was partially supported by a grant from the grants from National Natural Science Foundation of China (62372441, U22A2034), in part by Guangdong Basic and Applied Basic Research Foundation (2023A1515030268), and in part by Shenzhen Science and Technology Program (Grant No. RCYX20231211090127030, JCYJ20220818101401003).

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[3] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[4] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 5835–5843.

[5] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine CNN for image super-resolution," *IEEE Transactions on Multimedia*, vol. 9210, no. c, pp. 1–1, 2020.

[6] C. Tian, R. Zhuge, Z. Wu, Y. Xu, W. Zuo, C. Chen, and C.-W. Lin, "Lightweight image super-resolution with enhanced cnn," *Knowledge-Based Systems*, vol. 205, p. 106235, 2020.

[7] D. Qiu, Y. Cheng, and X. Wang, "Dual u-net residual networks for cardiac magnetic resonance images super-resolution," *Computer Methods and Programs in Biomedicine*, vol. 218, p. 106707, 2022.

[8] S. I. Young, Y. Balbastre, B. Fischl, P. Golland, and J. E. Iglesias, "Fully convolutional slice-to-volume reconstruction for single-stack mri," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 535–11 545.

[9] D. Mahapatra, B. Bozorgtabar, and R. Garnavi, "Image super-resolution using progressive generative adversarial networks for medical image analysis," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 30–39, 2019.

[10] W. Zhang, D. Yang, C. Y. Cheung, and H. Chen, "Frequency-Aware Inverse-Consistent Deep Learning for OCT-Angiogram Super-Resolution," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022, Proceedings, Part II*, 2022, pp. 645–655.

[11] B. Goyal, D. C. Lepcha, A. Dogra, and S.-H. Wang, "A weighted least squares optimisation strategy for medical image super resolution via multiscale convolutional neural networks for healthcare applications," *Complex & Intelligent Systems*, pp. 1–16, 2021.

[12] H. Lin, J. Zou, K. Wang, Y. Feng, C. Xu, J. Lyu, and J. Qin, "Dual-space high-frequency learning for transformer-based mri super-resolution," *Computer Methods and Programs in Biomedicine*, vol. 250, p. 108165, 2024.

[13] J. Li, H. Yang, Q. Yi, M. Lu, J. Shi, and T. Zeng, "High-frequency modulated transformer for multi-contrast mri super-resolution," *IEEE Transactions on Medical Imaging*, 2025.

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*, 2021, pp. 1833–1844.

[15] W. Huang, X. Liao, H. Chen, Y. Hu, W. Jia, and Q. Wang, "Deep local-to-global feature learning for medical image super-resolution," *Computerized Medical Imaging and Graphics*, vol. 115, p. 102374, 2024.

[16] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022*, 2022, pp. 456–465.

[17] C. Fang, D. Zhang, L. Wang, Y. Zhang, L. Cheng, and J. Han, "Cross-modality high-frequency transformer for mr image super-resolution," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1584–1592.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[19] D. Han, Z. Wang, Z. Xia, Y. Han, Y. Pu, C. Ge, J. Song, S. Song, B. Zheng, and G. Huang, "Demystify mamba in vision: A linear attention perspective," *Advances in neural information processing systems*, vol. 37, pp. 127 181–127 203, 2024.

[20] S. Li, H. Singh, and A. Grover, "Mamba-nd: Selective state space modeling for multi-dimensional data," in *European Conference on Computer Vision*. Springer, 2024, pp. 75–92.

[21] N. Muca Cirone, A. Orvieto, B. Walker, C. Salvi, and T. Lyons, "Theoretical foundations of deep selective state-space models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 127 226–127 272, 2024.

[22] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *European conference on computer vision*. Springer, 2024, pp. 222–241.

[23] X. Di, L. Peng, P. Xia, W. Li, R. Pei, Y. Cao, Y. Wang, and Z.-J. Zha, "Qmambabsr: Burst image super-resolution with query state space model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 080–23 090.

[24] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, and C.-W. Lin, "Frequency-assisted mamba for remote sensing image super-resolution," *IEEE Transactions on Multimedia*, 2024.

[25] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight Bimodal Network for Single-Image Super-Resolution via Symmetric CNN and Recursive Transformer," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 913–919, 2022.

[26] M. Li, Y. Chen, Z. Ji, K. Xie, S. Yuan, Q. Chen, and S. Li, "Image projection network: 3D to 2D image segmentation in OCTA images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3343–3354, 2020.

[27] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[28] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," *MedRxiv*, 2020.

[29] B. Puccio, J. P. Pooley, J. S. Pellman, E. C. Taverna, and R. C. Craddock, "The preprocessed connectomes project repository of manually corrected skull-stripped t1-weighted anatomical MRI data," *Gigascience*, vol. 5, no. 1, pp. s13 742–016, 2016.

[30] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 1646–1654.

[31] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*, 2014, pp. 111–126.

[32] Z. Xu, Y. Gao, X. Chen, K. Lin, L. Liu, and Y.-C. Chen, "Axial super-resolution in optical coherence tomography images via spectrum-based self-supervised training," *IEEE Transactions on Computational Imaging*, 2025.

[33] B. Yao, L. Jin, J. Hu, Y. Liu, Y. Yan, Q. Li, and Y. Lu, "Pscat: a lightweight transformer for simultaneous denoising and super-resolution of oct images," *Biomedical Optics Express*, vol. 15, no. 5, pp. 2958–2976, 2024.

[34] M. Lerendegui, K. Riemer, G. Papageorgiou, B. Wang, L. Arthur, A. Chavignon, T. Zhang, O. Couture, P. Huang, M. Ashikuzzaman *et al.*, "Ultra-sr challenge: Assessment of ultrasound localization and tracking algorithms for super-resolution imaging," *IEEE transactions on medical imaging*, vol. 43, no. 8, pp. 2970–2987, 2024.

[35] F. Gao, B. Li, L. Chen, X. Wei, Z. Shang, and C. Liu, "Ultrasound image super-resolution reconstruction based on semi-supervised cyclegan," *Ultrasonics*, vol. 137, p. 107177, 2024.

[36] H. Yuan, Q. Sun, Z. Wang, X. Fu, C. Ji, Y. Wang, B. Jin, and J. Li, "Dg-mamba: Robust and efficient dynamic graph structure learning with selective state space models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 21, 2025, pp. 22 272–22 280.

[37] H. Wang, Y. Chen, W. Chen, H. Xu, H. Zhao, B. Sheng, H. Fu, G. Yang, and L. Zhu, "Serp-mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model," *IEEE Transactions on Medical Imaging*, 2025.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[39] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016, pp. 1874–1883.